

Automatisierte Verschlagwortung und Zuordnung von Büchern in eine Systematik

Philipp Pollack

Masterarbeit

Beginn der Arbeit:	01. Juli 2016
Abgabe der Arbeit:	02. Januar 2017
Gutachter:	Prof. Dr. Stefan Conrad Prof. Dr. Martin Mauve

Erklärung

Hiermit versichere ich, dass ich diese Masterarbeit selbstständig verfasst habe. Ich habe dazu keine anderen als die angegebenen Quellen und Hilfsmittel verwendet.

Düsseldorf, den 02. Januar 2017

Philipp Pollack

Zusammenfassung

Der Bestand der Zentralbibliothek (ZB) des Forschungszentrums Jülich beinhaltet über 127.000 Bücher und 147.000 E-Bücher. Bei diesem Umfang an Büchern ist eine Sacherschließung unumgänglich, um eine gute Durchsuchbarkeit zu ermöglichen. Die Sacherschließung umfasst die Zuweisung von Systemstellen sowie die Verschlagwortung und erfolgt momentan durch die fünf Sachbearbeiter der ZB. Die Systematik ordnet die Bücher in Sachgebiete ein. Dabei gibt es Systemstellen in verschiedener Granularität. Eine grobe Einteilung ist z.B. Chemie, Informatik, Medizin, usw., wohingegen bei einer feineren Einteilung die Obergebiete nochmals aufgeteilt werden. Für die Verschlagwortung kommt in der ZB aktuell ein Katalog mit über 7.700 Schlagwörtern aus allen Bereichen zum Einsatz. Ein Schlagwort beschreibt im Allgemeinen einen Aspekt des Buches und muss nicht selber wörtlich in diesem vorkommen.

Im Rahmen dieser Arbeit soll mittels Machine Learning sowohl eine automatische Einordnung von neuen Büchern sowie E-Büchern in eine Systematik als auch deren Verschlagwortung erfolgen. Das Ziel ist, dass die Sachbearbeiter gegebenenfalls nur noch die Korrektheit der automatisch erzeugten Vorschläge prüfen müssen. Dazu werden verschiedene Verfahren bzw. Klassifikatoren betrachtet. Neben dem bereits bewährten k -Nächste-Nachbarn Klassifikator und einem neuronalen Netz, wird im Rahmen dieser Arbeit ein neues graphenbasiertes Verfahren namens BookGraph entwickelt.

Zu beiden Aufgabenteilen wird eine Bewertung der Ergebnisse und ein Vergleich der verschiedenen Klassifikatoren durchgeführt. Die Bewertung der Verfahren wird zum einen auf Basis der bereits katalogisierten Bücher durchgeführt und zum anderen durch eine manuelle Auswertung durch die Sachbearbeiter der ZB. Bei allen getesteten Szenarien konnte sich das BookGraph Verfahren von den beiden anderen abheben.

Im ersten Kapitel wird die Masterarbeit motiviert und die wichtigsten Begriffe erläutert. In Kapitel 2 werden die Grundlagen der Arbeit aufgegriffen und Hintergrundinformationen über die Metadaten der Bücher gegeben. Das Kapitel 3 stellt die drei in dieser Arbeit behandelten Verfahren vor und erläutert deren Funktionsweise. Daraufhin werden die Ergebnisse der Verfahren in Kapitel 4 analysiert und die Leistung miteinander verglichen. Eine Beispielanwendung zur Demonstration des zukünftigen Einsatzes ist Inhalt des Kapitels 5. Abschließend werden in Kapitel 6 die Ergebnisse der Arbeit zusammengefasst und die Vor- und Nachteile der verschiedenen Verfahren besprochen. Zusätzlich wird ein kurzer Ausblick gegeben.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Motivation	1
1.2	Ähnliche Arbeiten	1
1.3	Begriffserklärung	2
1.3.1	Buch / E-Buch	2
1.3.2	Sacherschließung	2
1.3.3	Multi-Label Klassifikation	3
1.3.4	Gold-Standard Werte	3
1.4	Zielsetzung der Masterarbeit	3
2	Grundlagen	5
2.1	Die Zentralbibliothek des Forschungszentrums Jülich	5
2.2	Bisherige Erfassung der Bücher	5
2.3	Metadaten	6
2.4	Verschlagwortung	7
2.5	Systematik	8
2.6	Word2Vec	8
3	Die Verfahren	11
3.1	BookGraph	11
3.1.1	Der Graph	11
3.1.2	Bestimmung von Systemstellen	13
3.1.3	Bestimmung von Schlagwörtern	14
3.2	k -Nächsten-Nachbarn	15
3.2.1	Feature-Vektoren	15
3.2.2	Abstandsmaß	16
3.3	Neuronales Netz	17
4	Bewertung der Verfahren	19
4.1	Testdaten	19

4.2	Benutzte Evaluationsmaße	21
4.2.1	Deckungsgrad	22
4.2.2	Hamming Loss	22
4.2.3	Accuracy, Precision und Recall	23
4.2.4	Treffer-sicherheit	24
4.2.5	Single-Label Betrachtung	24
4.3	Auswertung für Systemstellen	25
4.3.1	BookGraph	25
4.3.2	k -Nächste-Nachbarn	29
4.3.3	Neuronales Netz	31
4.4	Auswertung für Schlagwörter	33
4.4.1	BookGraph	34
4.4.2	k -Nächste-Nachbarn	36
4.4.3	Neuronales Netz	38
4.5	Anwendung auf Testdaten	40
4.6	Auswertung der Treffer-sicherheit	42
4.7	Auswertung für Single-Label	44
4.8	Evaluation des Praxiseinsatzes durch Sachbearbeiter	46
5	Programm zur Demonstration der Verfahren	51
5.1	Programmstrukturierung	51
5.2	Workflow	51
5.3	Grafische Benutzeroberfläche	52
6	Fazit und Ausblick	55
6.1	Zusammenfassung und Fazit	55
6.2	Ausblick	56
6.2.1	Weiterentwicklung der Verfahren	56
6.2.2	Produktiveinsatz	56
A	Auszug der Systemstellen	65
B	Auszug der Schlagwörter	67
C	Einzelergebnisse Systemstellen Auswertung	69
D	Einzelergebnisse Schlagwörter Auswertung	75

E	Einzelergebnisse Single-Label Auswertung	79
F	Einzelergebnisse Testdaten	81

1 Einleitung

Dieses Kapitel behandelt zunächst die Motivation dieser Arbeit und Erläuterung der Problemstellung. Darauf folgt eine Erklärung der wichtigsten Begriffe und eine Zusammenfassung der Zielsetzung dieser Masterarbeit.

1.1 Motivation

Die Zentralbibliothek (ZB) des Forschungszentrums Jülich betreibt eine Datenbank mit allen Publikationen, die sich im Besitz der ZB befinden oder von Dritten lizenziert sind. Diese Datenbank ist die zentrale Stelle, wenn die Wissenschaftlerinnen und Wissenschaftler des Zentrums nach passender Literatur für ihre Arbeit suchen. Somit ist es von essentieller Wichtigkeit, dass der Bestand auf einfache und effiziente Weise durchsucht werden kann.

Ein wichtiges Mittel, um dies sicher zu stellen, ist die Sacherschließung der Bücher. Dabei werden jedem Buch Systemstellen sowie Schlagwörter zugewiesen, über die in der Datenbank gesucht werden kann. Momentan erfolgt die Sacherschließung durch die Sachbearbeiter der Zentralbibliothek, die in Handarbeit jedes neue Buch mustern und entsprechend die Systemstellen und Schlagwörter vergeben. Da dies einen hohen zeitlichen Aufwand erfordert, werden aktuell nur gedruckte Bücher sacherschlossen. Elektronische Bücher hingegen werden nicht bearbeitet und sind dadurch zum Beispiel bei einer Suche mittels der Schlagwörter nicht auffindbar.

Im Rahmen dieser Arbeit soll deswegen ein automatisches System zur Entlastung der Sachbearbeiter bei der Sacherschließung entwickelt werden.

1.2 Ähnliche Arbeiten

Mit der Thematik der automatischen Sacherschließung von Büchern befassen sich mittlerweile einige Bibliotheken. Da allerdings jede Bibliothek intern einen eigenen Aufbau hat und eine andere Methodik zur Erfassung der Bücher nutzt, gibt es kein einheitliches Framework für die automatische Sacherschließung.

2011 fand in der Deutschen Nationalbibliothek im Rahmen des Projektes PETRUS ein Treffen von verschiedenen Institutionen für den Erfahrungsaustausch auf diesem Gebiet statt¹. Bei diesem Workshop wurden eine Reihe von Ansätzen vorgestellt, allerdings ohne tiefere technische Details zu nennen.

Es zeigen sich aber zwei grundsätzlich verschiedene Ansätze. Zum einen werden kommerzielle Fremd-Produkte eingesetzt und zum anderen kommen Eigenentwicklungen zum Einsatz. Beiden gemein ist, dass meist ein Vorschlagssystem gewählt wird, sodass keine vollautomatische Erschließung angewandt wird. Statt dessen wird ein halbautomatisches System benutzt, bei der ein Sachbearbeiter schlussendlich eine Auswahl trifft.

1.3 Begriffserklärung

Dieser Abschnitt erläutert wichtige Begriffe, die im Folgenden vorkommen oder als Wissensgrundlage für das Verständnis notwendig sind.

1.3.1 Buch / E-Buch

Diese Arbeit beschäftigt sich mit Büchern und E-Büchern. Hier soll kurz der Unterschied zwischen beiden Begriffen beleuchtet werden. Der Begriff Bücher bezeichnet im Rahmen dieser Arbeit kontextabhängig ein Buch in gebundener bzw. gedruckter Form oder den Überbegriff für gedruckte und elektronische Bücher. Im Gegensatz dazu ist ein E-Buch ausschließlich in digitaler Form, z. B. als PDF-Datei, verfügbar.

1.3.2 Sacherschließung

Das Ziel vieler Bibliotheken, Museen und anderer Einrichtungen ist es ihren Bestand nicht nur zu erhalten, sondern auch nutzbar und zugänglich zu machen [Ebe11]. Dazu wird das Gesammelte untersucht und in verschiedene Ordnungen sortiert. Jede Ordnung stellt dabei einen anderen Zugang zu dem Bestand dar, beispielsweise eine chronologische Sortierung oder eine geographische Zuordnung.

In der ZB werden bei der Sacherschließung hauptsächlich zwei Ordnungen genutzt. Die Sachbearbeiter der ZB weisen zum einen den Büchern Schlagwörter zu und zum anderen ordnen sie die Bücher in eine Systematik nach Themengebieten ein.

¹http://www.dnb.de/DE/Wir/Projekte/Archiv/petrus_workshop.html

1.3.3 Multi-Label Klassifikation

Bei der klassischen Single-Label Klassifikation wird jedem Objekt aus der Menge der zu klassifizierenden Objekte genau ein Label zugeordnet. Stattdessen kann einem Objekt bei der Multi-Label Klassifikation auch mehrere Label zugeordnet werden. Ein Label bezeichnet dabei ein dem Objekt zugeordnetes Merkmal.

Da die im Rahmen dieser Arbeit betrachteten Bücher mehrere Schlagwörter bzw. Systemstellen haben können, handelt es sich um eine Multi-Label Klassifikation. Jedes Schlagwort bzw. jede Systemstelle stellt dabei ein Label dar und die Bücher die zu klassifizierenden Objekte.

Die im weiteren Verlauf benutzten Verfahren sind, mit Ausnahme von Kapitel 4.7, Multi-Label Klassifikatoren, da ansonsten die Problemstellung nicht erfüllt werden kann.

1.3.4 Gold-Standard Werte

Gold-Standard (GS) Werte werden bei Klassifikationen und ähnlichem als Vorgaben für die Ergebnisse genommen und gelten damit in diesem Kontext als richtig. Meist werden GS-Werte durch Menschen, z. B. mittels Crowdsourcing² ermittelt bzw. festgelegt. Aber auch eine maschinelle Erzeugung ist möglich.

In dieser Arbeit entsprechen die GS-Werte den Schlagwörtern bzw. Systemstellen, die bereits den Büchern im Bestand der Zentralbibliothek zugeordnet sind. Diese wurden durch Sachbearbeiter ermittelt und werden im Rahmen der Arbeit als korrekt angenommen.

1.4 Zielsetzung der Masterarbeit

Die Mitarbeiter der Zentralbibliothek des Forschungszentrum sollen bei der Sacherschließung neuer Bücher unterstützt werden.. Daher soll in dieser Masterarbeit ein Programm zur automatischen Bestimmung von Systemstellen sowie Schlagwörtern für Bücher und E-Bücher erstellt werden.

Für das zugrunde liegende Verfahren sollen verschiedene Ansätze betrachtet und miteinander verglichen werden. Aus der Struktur der Daten und der Zielsetzung ergeben sich verschiedene Szenarien, die gesondert betrachtet werden müssen. Zum einen gibt es im Bestand der Zentralbibliothek sowohl gedruckte als auch elektronische Bücher, die teils verschiedene Metadaten besitzen. Zum anderen sollte die automatische Verschlagwortung und Bestimmung von Systemstellen für beide Bucharten funktionieren.

²<http://www.gruenderszene.de/lexikon/begriffe/crowdsourcing>

2 Grundlagen

In diesem Kapitel werden zunächst einige Grundlagen näher erklärt, sowie die Rahmengegebenheiten aufgeführt. Dazu wird kurz die Zentralbibliothek vorgestellt und dann näher auf den Buchbestand und die Metadaten der Bücher eingegangen.

2.1 Die Zentralbibliothek des Forschungszentrums Jülich

Die Zentralbibliothek (ZB)¹ ist ein wichtiger Dienstleister für die 5.500 Mitarbeiterinnen und Mitarbeiter des Forschungszentrums Jülich. Sie bietet viele Dienste rund um die Beschaffung und Veröffentlichung von Informationen. So ist die Zentralbibliothek in die Bereiche Sprachendienst, Wissenschaftliche Dienste und Literaturerwerbung aufgeteilt. Daneben existieren zusätzlich noch Stabsstellen für besondere Aufgabenbereiche. Ein wichtiges Arbeits- und Informationsmittel für die Mitarbeiter bildet der umfassende Bestand an Publikationen. Die Zentralbibliothek verfügt über 127.000 gedruckte Bücher, 147.000 elektronische Bücher (E-Books) und 70.000 Zeitschriften. Dabei sind die meisten der elektronischen Bücher nicht im direkten Besitz der ZB, sondern werden mittels Abonnements bzw. Verträgen von den Verlagen zur Verfügung gestellt. Alle Publikationen können über das Portal JuLib eXtended², einer Onlinesuche mit erweiterten Funktionen, durchsucht und bei Bedarf direkt ausgeliehen werden.

2.2 Bisherige Erfassung der Bücher

Bücher werden in der Zentralbibliothek auf Anfrage der Mitarbeiter oder einem vermutlichen hohen Interesse an der Publikation bestellt. Seltener kommt es zu einer hohen Anzahl gleichzeitig neu bestellter Bücher. Nach Erhalt der Bücher werden diese in der Dokumentlieferung digital erfasst und im System eingetragen. Dabei wird auch das Inhaltsverzeichnis mittels Optical Character Recognition (OCR) erfasst und gespeichert. Als nächstes werden die Bücher nach Themengebieten aufgeteilt und an einen der fünf

¹http://www.fz-juelich.de/zb/DE/Home/home_node.html

²<http://julib.fz-juelich.de>

Sachbearbeiter weitergegeben. Dieser verschafft sich einen Überblick über das Buch und vergibt dann eine oder mehrere Systemstellen und Schlagwörter. Zusammengefasst werden die Metadaten des Buches in einem MARC [LML⁺99]-Eintrag im Bibliothekssystem. Schließlich werden die MARC-Einträge genutzt, um den Index der Suchmaschine JuLib der Zentralbibliothek zu speisen.

Bei elektronischen Büchern hingegen ist es eher unüblich sie einzeln zu kaufen. Meist werden diese als Pakete mit einem Umfang von mehreren tausend Büchern erworben, oder stehen per Vertrag mit einem Verlag zur Verfügung. Dabei erfolgt die Auslieferung bzw. Bereitstellung der E-Bücher direkt als entsprechende MARC-Dateien, welche so einfach in den Bibliothekskatalog übernommen werden. Im Gegensatz zu gedruckten Büchern erfolgt aufgrund des großen Umfangs keine Zuweisung von Schlagwörtern oder Systemstellen durch die Sachbearbeiter. Allerdings besitzen die meisten E-Bücher bereits vom Verlag vergebene Schlagwörter sowie Systemstellen nach dem Katalog der Library-of-Congress [Lib16].

2.3 Metadaten

Feld	Beschreibung
id	Interner Primärschlüssel
author2	Beteiligte Autoren
isbn	ISBN
language	Verwendete Sprachen
fulltext	Volltext (meist nur E-Books)
title_full	Kompletter Titel
classification	Systemstelle
subject_headings_zb	Schlagwörter der ZB
table_of_contents	Inhaltsverzeichnis
edition	Edition
publisher	Herausgeber

Tabelle 1: Wichtigste Metadaten einer Publikation

Um eine gute Durchsuchbarkeit in JuLib zu gewährleisten, werden zu jeder Publikation diverse Metadaten erhoben. Gespeichert werden diese in MARC-Dateien bzw. auf einem Apache Solr Server. Die wichtigsten Metadaten sind in Tabelle 1 aufgeführt. Jeder Eintrag in JuLib besitzt einen Primärschlüssel, um die Publikation eindeutig im Bestand identifizieren zu können. Auch besitzt jeder Eintrag allgemeine Daten zu beispielsweise Titel, Autor, Sprache, Edition und Herausgeber. Dazu kommen noch Metadaten, die nicht

ID	250494
Titel	Parallele und verteilte Anwendungen in Java
Autor	Oechsle, Rainer
Sprache	deutsch
Systemstellen	DOR - Programming languages - JAVA, C#, Mono, .NET
Schlagwörter	parallel algorithm, parallel processing, Java (programming language), network, Internet

Tabelle 2: Beispiel für ein Buch

bei allen Büchern vorhanden sind. Dies sind vor allem: die Systemstellen, die Schlagwörter, das Inhaltsverzeichnis oder auch der komplette Volltext. Ein Buch kann auch mehrere Systemstellen oder Schlagwörter besitzen, welche keine feste Reihenfolge haben. Ein Beispiel für ein Buch ist in Tabelle 2 angegeben. Das Buch besitzt eine Systemstelle und fünf Schlagwörter. Dieses Beispiel wird in den nachfolgenden Kapiteln für weitere Veranschaulichungen genutzt.

2.4 Verschlagwortung

Bei einer Verschlagwortung werden Deskriptoren einem Dokument zugeordnet, um die darin enthaltenen Sachverhalte zu erschließen. Dazu wird meist auf einen vorgefertigten Schlagwortkatalog zurückgegriffen, aber auch eine freie Verschlagwortung ist möglich. Im Allgemeinen ist ein Schlagwort ein Begriff, der den Inhalt eines Dokuments beschreibt und zur Durchführung einer systematischen Recherche z.B. in einem Bibliothekskatalog dient. Ein Schlagwort muss auch nicht selbst im Dokument vorkommen, sondern beschreibt das Dokument nur [Def16a].

Eine Verschlagwortung kann mit Hilfe mehrerer verschiedener Methoden durchgeführt werden: manuell, automatisch oder computergestützt / halbautomatisch. Die manuelle Verschlagwortung erfolgt von einem Experten, der anhand von Regeln Schlagwörter vergibt. Wohingegen eine automatische Verschlagwortung von einem Computerprogramm durchgeführt wird, welches selbstständig ohne menschliches Zutun Schlagwörter auswählt. Eine Kombination beider Ansätze stellt die computergestützte oder auch halbautomatische Verschlagwortung dar. Ein Computerprogramm schlägt mögliche Schlagwörter vor, aber ein Mensch wählt die Schlagwörter schließlich aus.

In der ZB erfolgt die Verschlagwortung manuell anhand eines festen Schlagwortkatalogs, der 7.730 Begriffe enthält (Stand Juni 2015). Aktuell machen dies fünf Sachbearbeiter, welche nach Fachbereich aufgeteilt, den Büchern Schlagwörter zuweisen. E-Books

werden allerdings noch nicht nach dem Schlagwortkatalog der ZB verschlagwortet. Diese besitzen meist nur Schlagwörter nach dem Katalog der Library-of-Congress, welche von den Verlagen mit jeweils eigenen Kriterien vergeben werden. Ein Auszug aus den Schlagwörtern der ZB hängt der Arbeit an (Anhang B).

2.5 Systematik

Eine Systematik ordnet die Dokumente eines Bibliotheksbestands in thematisch passende wissenschaftliche Teilgebiete ein. Ein Fachgebiet wird dabei Systemstelle genannt. Die Systemstellen sind hierarchisch aufgebaut, so teilt sich ein allgemeines Teilgebiet in immer mehr speziellere Teilgebiete auf [Def16b]. In Tabelle 3 ist ein Beispiel aus der Systematik der ZB dargestellt.

Systemstelle	Fachgebiet
C	Chemistry
CE	Organic chemistry
CEX	Special organic reactions
CEXG	Organic reactions - oxidation reactions

Tabelle 3: Beispiel aus der Systematik

In der ZB kommt eine fünfstufige Systematik zur Anwendung. Insgesamt gibt es 1.931 Systemstellen. Diese gliedern sich in 13 einstellige, acht zweistellige, 1.028 dreistellige, 877 vierstellige und fünf fünfstelligen Systemstellen. Ein Auszug aller Systemstellen ist im Anhang A einzusehen. Die ein- und zweistelligen Systemstellen werden in der ZB nicht mehr an neue Bücher verteilt, da diese eine zu ungenaue Einteilung ergeben. Da viele Bücher aus dem Bestand allerdings schon diese kurzen Systemstellen zugewiesen bekommen haben, werden sie im Rahmen dieser Arbeit nicht ausgenommen.

Die Systematik wird auch für die Aufstellung der Bücher im Lesesaal genutzt, dort werden allerdings der Übersicht halber Teilgebiete zusammengefasst, um eine Fragmentierung zu verhindern.

2.6 Word2Vec

Word2Vec ist ein Modell zur Erzeugung von Wortembeddings [MCCD13]. Eine Wortembedding ist die Repräsentation eines Wortes als numerischer Vektor. Ein semantischer Vergleich zwischen Wörtern ist dann mittels Vergleich der Vektoren möglich. In Word2Vec

liegen die Vektoren von semantisch ähnlichen Wörtern dichter beisammen, als bei unähnlicheren Wörtern. Die genaue Funktionsweise von Word2Vec soll an dieser Stelle nicht erläutert werden, sondern es sei auf [MSC⁺13] verwiesen.

Die Vektoren zu den Wörtern müssen zunächst anhand eines Trainingskorpus gelernt werden. Dazu werden in dieser Arbeit die deutsche als auch die englische Wikipedia, mit Stand zum 29.11.2016, genutzt. Englisch und Deutsch werden basierend auf dem Schwerpunkt des Buchbestandes der ZB ausgewählt. Es existieren zwar auch anderssprachige Bücher im Bestand, diese stellen allerdings eine Randerscheinung dar und sind für den täglichen Gebrauch zweitrangig.

Für beide Korpora werden dann die Vektoren erzeugt, dabei wird eine Dimensionalität von 200 gewählt. Zur Erzeugung der Vektoren wurde die Python Bibliothek *gensim* genutzt, welche entsprechende Methoden zur Verfügung stellt [ŘS10].

3 Die Verfahren

In diesem Kapitel werden die verschiedenen Verfahren zur automatischen Klassifikation und Verschlagwortung vorgestellt, die in dieser Masterarbeit genutzt werden. Das erste Verfahren ist ein im Rahmen dieser Arbeit neu entwickeltes Verfahren, welches als zentralen Bestandteil einen Graphen nutzt. Im Folgenden wird dieses BookGraph bezeichnet. Das zweite Verfahren nutzt den verbreiteten k -Nächste-Nachbarn Klassifikator zur Bestimmung der Systemstellen und Schlagwörter. Das dritte und letzte Verfahren ermittelt die Vorschläge mit Hilfe eines neuronalen Netzes.

3.1 BookGraph

In diesem Abschnitt wird das eigens im Rahmen dieser Masterarbeit entwickelte Verfahren zur Bestimmung von Systemstellen sowie Schlagwörtern vorgestellt. Das Verfahren basiert auf einem Graphen, welcher für die Bestimmung der Systemstellen bzw. Schlagwörter ausgewertet wird.

3.1.1 Der Graph

Hauptbestandteil des entwickelten Verfahrens zur Bestimmung der Systemstellen und Schlagwörter bildet ein Graph [AL99]. Ein Graph $G = (V, E)$ besteht aus einer Menge an Knoten V , welche über Kanten $E \subseteq V \times V$ miteinander verbunden sind. In dem hier benutzten Graphen repräsentieren die Knoten bekannte Informationen aus Büchern und die Kanten stellen Beziehungen der Knoten untereinander dar. Haben zum Beispiel zwei Autoren zusammen an einem Buch geschrieben, wird für jeden Autor ein Knoten angelegt und diese beiden mit einer Kante verbunden. Analog wird für andere bekannte Informationen vorgegangen. Mögliche Informationsarten sind Autor, Schlagwort und Systemstelle.

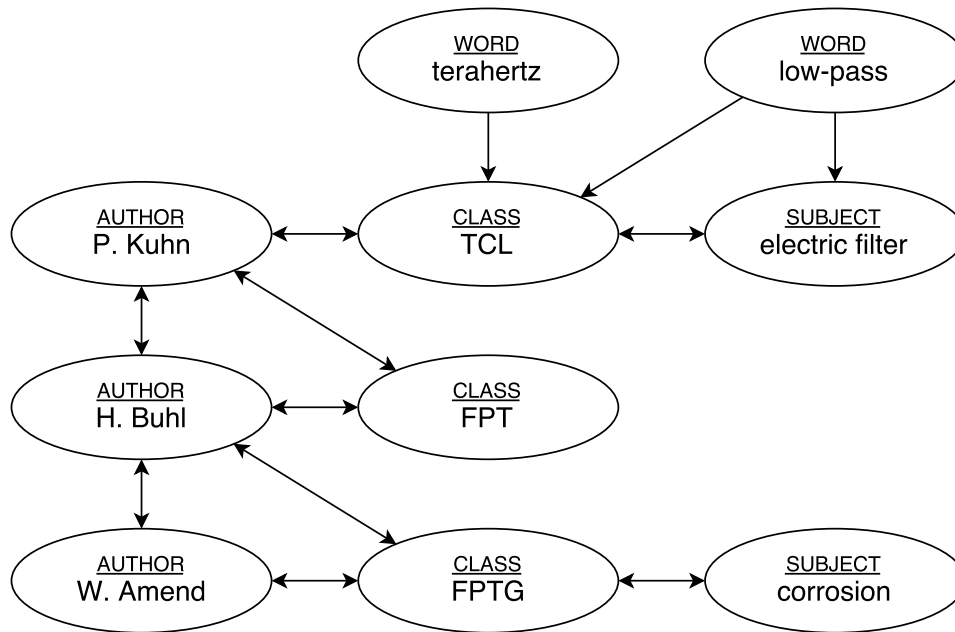


Abbildung 1: Auszug aus einem Graphen

Knoten

Jeder Knoten besitzt einen zugeordneten Knotentypen, um die Informationsart festzuhalten. Die möglichen Knotentypen sind: *Subject* (Schlagwort), *Class* (Systemstelle), *Author* (Autor) und *Word* (Wort des Inhaltsverzeichnis). Die Information selber wird in einer Eigenschaft *Namen* gespeichert. Knotentyp und Name sind zusammen eindeutig und können zur Knotenidentifikation genutzt werden. Ein Beispielauszug aus einem Graphen ist in Abbildung 1 dargestellt.

Damit gehört zu einem Knoten ein Name, ein Typ, sowie Menge an eingehenden und ausgehenden Kanten. Zusätzlich verfügt ein Knoten über eine Menge an Büchern, in denen die gespeicherte Information vorkommt. Der Knotentyp *Word* nimmt eine Sonderrolle ein. Sofern ein Buch über ein digitalisiertes Inhaltsverzeichnis verfügt, wird dieses zunächst in die einzelnen Wörter zerlegt. Aus den Wörtern werden dann die Stoppwörter herausgefiltert und die Übriggebliebenen als Knoten des Typs *Word* in den Graphen eingefügt. Dabei werden keine Kanten zwischen zwei *Word*-Knoten eingefügt sondern nur Kanten von *Word*-Knoten zu Knoten der anderen Typen.

Kanten

Kanten verbinden wie bereits erwähnt zwei Knoten miteinander, deren Information zusammen in einem Buch vorkommt, dabei ist der Knotentyp unerheblich. Grundsätzlich sind Kanten gerichtet, sodass für ein gemeinsames Auftreten in in einem Buch zwei Kanten angelegt werden müssen. Als Kantengewicht wird die Häufigkeit des gemeinsamen Vorkommens genommen. Dadurch, dass nur Kanten zwischen Knoten, deren Informationen gemeinsam in einem Buch vorkommen, erstellt werden, kann es vorkommen, dass der resultierende Graph nicht zusammenhängend ist, sondern aus mehreren unabhängigen Teilgraphen besteht.

3.1.2 Bestimmung von Systemstellen

Ist ein Graph nach Abschnitt 3.1.1 gegeben, können mit seiner Hilfe die Systemstellen für ein bis dahin unbekanntes Buch bestimmt werden. Die Bestimmung erfolgt verfahrensbedingt nicht absolut, sondern in Relation zu Ähnlichkeiten, die bereits im Graphen existieren. Dies bedeutet, je mehr Bücher als Trainingsdaten in den Graphen eingefügt werden, umso besser können Zusammenhänge zwischen den einzelnen Informationen bestimmt werden.

Sollen zu einem Buch nun mittels des Graphen die Systemstellen bestimmt werden, werden, als erster Schritt, die Autoren des Buches betrachtet. Für jeden Autor wird im Graphen nachgeschaut, ob dieser dort als Knoten vertreten ist. Falls dies zutrifft, werden alle Systemstellen in die Vorschlagliste aufgenommen, die über eine einzelne Kante erreicht werden können. Im Falle, dass kein Autor des Buches im Graphen vorhanden ist, werden die Wörter des Inhaltsverzeichnisses betrachtet. Analog zu den Autoren werden für jedes auch im Graphen vorkommende Wort, die Systemstellen, die über eine direkte Kante erreicht werden können, in die Vorschlagliste aufgenommen.

Um relevante von eher nicht relevanten Vorschlägen zu trennen, werden sowohl bei den Autoren als auch bei den Wörtern nur Kanten zu Systemstellen betrachtet, welche über einen bestimmten Schwellwert t an lokaler relativer Häufigkeit besitzen. Dazu wird das summierte Kantengewicht aller Kanten eines Knotens bestimmt und ins Verhältnis zum einzelnen Kantengewicht gesetzt. Optional kann ein dynamischer Schwellwert benutzt werden, welcher bei weniger als zehn Vorschlägen die Schwelle graduell um 0.01 senkt, bis mehr als zehn Vorschläge zusammen kommen oder die Schwelle auf 0 gesunken ist. Der Algorithmus zur Bestimmung einer Vorschlagliste für die Systemstellen ist in Abbildung 2 zu sehen.

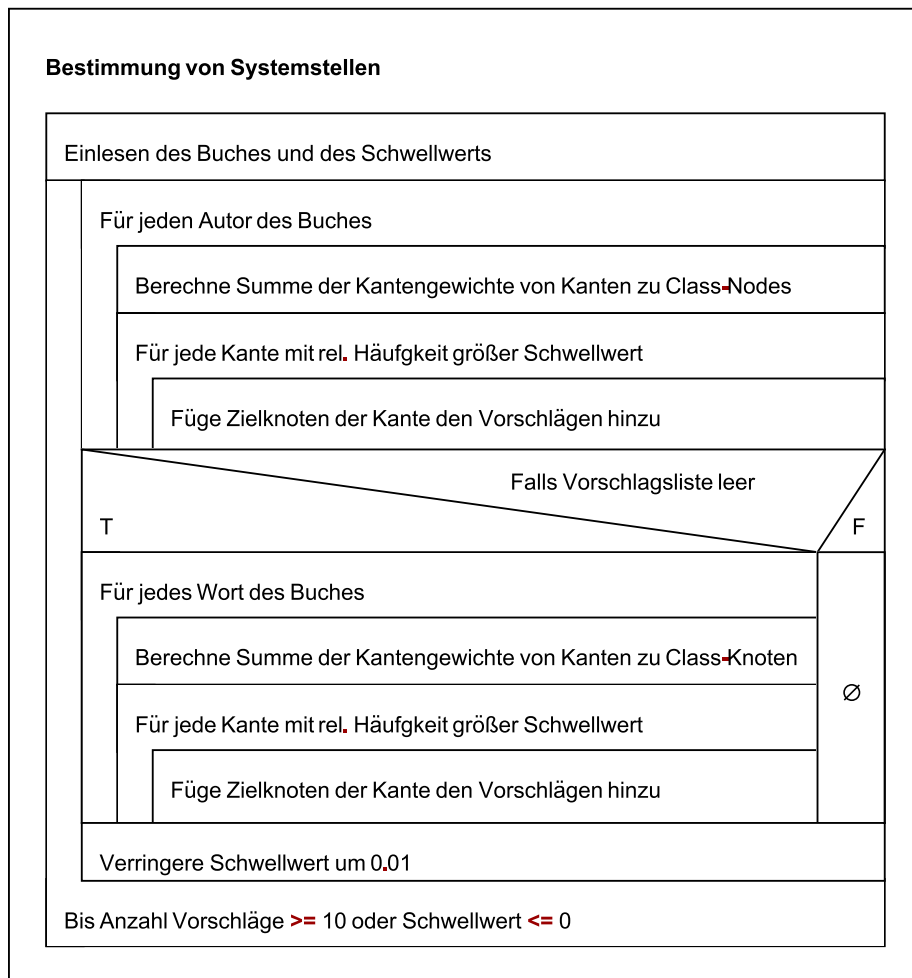


Abbildung 2: Algorithmus zur Bestimmung der Systemstellen

3.1.3 Bestimmung von Schlagwörtern

Die Bestimmung der Schlagwörter funktioniert identisch zu der Bestimmung der Systemstellen, mit dem Unterschied, dass bei den Schlagwörtern auch Kanten zweiten Grades betrachtet werden. Dass heißt es werden auch Knoten für die Vorschlagsliste ausgewählt, die zwei Kanten von den ursprünglichen Autor- bzw. Wort-Knoten entfernt sind. Dabei wird auch für diese Knoten der Schwellwert beachtet, indem die Wahrscheinlichkeiten am Pfad entlang multipliziert werden. Das Betrachten von Kanten zweiten Grades ist bei der Bestimmung von Schlagwörtern sinnvoll, da es oft zueinander ähnliche Schlagwörter gibt, die eine ähnliche Bedeutung haben oder oft zusammen auftreten. Dagegen ist dies bei den Systemstellen nicht sehr häufig, sodass es nicht sinnvoll ist, von

einer gefunden Systemstellen die benachbarten Systemstellen zu betrachten.

Sowohl beim Vorgehen für die Bestimmung der Schlagwörter als auch bei dem für Systemstellen wurden während der Entwicklung mehrere Möglichkeiten ausprobiert. Die hier vorgestellten Regeln zur Bestimmung stellen den im Laufe der Arbeit verbesserten Stand dar.

3.2 k -Nächsten-Nachbarn

Das zweite Verfahren zur Bestimmung der Systemstellen und Schlagwörter, welches betrachtet wird, ist der k -Nächste-Nachbarn (KNN) Klassifikator [CH06]. KNN ist ein gebräuchliches Verfahren im Bereich des Machine Learnings.

Jedem Buch wird ein Feature-Vektor (siehe Abschnitt 3.2.1) mit n Einträgen und damit eine Position im n -dimensionalen Raum zugeordnet. Anhand des Abstands dieser Positionen wird dann die Ähnlichkeit zwischen den Büchern bestimmt. Dabei werden die k Nachbarn mit den kleinsten Abständen gewählt. Auf Basis dieser k Bücher wird zur Bestimmung der Systemstellen bzw. Schlagwörter eine Liste der vorkommenden Systemstellen bzw. Schlagwörtern und deren relative Häufigkeiten erstellt. Die sortierte absteigende Liste bildet die Vorschlagsliste des KNN Klassifikators.

In Abbildung 3 ist ein Beispiel für den zweidimensionalen Fall dargestellt, in diesem soll ein neues Element in eine von zwei Klassen eingeordnet werden.

Für den KNN Klassifikator sind folglich ein Verfahren zur Erzeugung der Feature-Vektoren sowie ein Maß zur Berechnung der Ähnlichkeit notwendig. Diese beiden Punkte werden im Folgenden beschrieben. Dabei wird unter anderem ein erster Versuch zur Erstellung der Feature-Vektoren vorgestellt, als auch die finale Version, welche in der weiteren Arbeit genutzt wird.

3.2.1 Feature-Vektoren

Jedes Buch besitzt einen Feature-Vektor als Repräsentanten. Jede Dimension des Vektors ist dabei ein anderes Feature, also ein Merkmal des Buches. Ein Beispiel für ein solches Feature wäre zum Beispiel das Erscheinungsjahr oder der Autor eines Buches. Die Features müssen numerisch erfassbar sein, um sie als Vektor darstellen zu können. So wird eine alleinig auf Zahlen basierende Darstellung der Bücher erreicht.

Die Auswahl der Features spielt dabei eine sehr wichtige Rolle, da nur durch eine gute Auswahl der Features eine hinreichend gute Vergleichbarkeit von Ähnlichkeiten und Unterschieden zwischen den Büchern möglich ist.

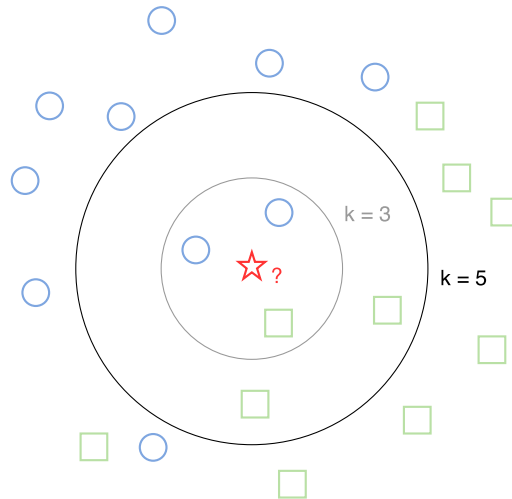


Abbildung 3: Beispiel mit verschiedenen k

Zunächst wurde als Feature-Vektor ein binäres Bag-of-Words Modell genutzt. Bei diesem Bag-of-Words Modell steht jede Dimension des Vektors für ein festes Wort und der boolesche Wert gibt das Vorhandensein des Wortes in dem betreffenden Buch an.

Als Basis für das Vokabular des Bag-of-Words Modell dienen bei den gedruckten Büchern die Inhaltsverzeichnisse der Bücher. Aus allen vorkommenden Wörtern in allen Inhaltsverzeichnissen wird das Vokabular des Bag-of-Words Modell gebildet, wobei in jedem Feature-Vektor eines Buches die Häufigkeiten des Inhaltsverzeichnisses des dazugehörigen Buches gespeichert werden. Somit kann eine Unterscheidung der Bücher auf Basis der im Inhaltsverzeichnis vorkommenden Wörter getroffen werden.

Diese erste Variante für die Erstellung der Feature-Vektoren wurde dann durch eine andere ersetzt. Die Feature-Vektoren werden nun mittels der Word2Vec Vektoren (siehe Abschnitt 2.6) gebildet. Zunächst werden für die Menge aller im Inhaltsverzeichnis eines Buches vorkommenden Wörter, ohne die Stoppwörter, die Word2Vec Vektoren bestimmt. Danach wird der Zentroid dieser Vektoren berechnet. Dieser dient als Repräsentant und damit als der Feature-Vektor des Buches.

3.2.2 Abstandsmaß

Mit Hilfe des Abstandsmaßes werden die Feature-Vektoren miteinander verglichen und damit die Ähnlichkeit von zwei Büchern zueinander bestimmt. Ein kleiner Abstand entspricht dabei einer größeren Ähnlichkeit. Mit wachsendem Abstand werden sich also zwei Bücher immer unähnlicher. Eines der gängigsten Maße ist der euklidische Abstand.

Für die Ähnlichkeitsbestimmung der Bücher wird allerdings die Manhattan-Metrik als

ein einfacheres Abstandsmaß gewählt, um die Berechnungszeit zu senken. Das Bag-of-Words aller Wörter der Inhaltsverzeichnisse weist über 700.000 Einträge auf, damit haben auch die Feature-Vektoren ebenso viele Einträge. Der Abstand zwischen zwei Feature-Vektoren $A = (a_1, \dots, a_n)$ und $B = (b_1, \dots, b_n)$ ist die Summe der absoluten Differenzen aller Dimensionen:

$$d(A, B) = \sum_{i=1}^n |a_i - b_i| \quad (3.1)$$

Diese Metrik wurde bzw. wird für beide vorgestellten Feature-Vektoren genutzt. Bei der ersten Version auf Basis der Bag-of-Word Modelle ist die Manhattan-Metrik nur die symmetrische Differenz der vorkommenden Wörter. Bei der finalen Variante wird die Metrik, wie in Formel 3.1 dargestellt, berechnet.

3.3 Neuronales Netz

Der dritte Klassifikator, der in dieser Arbeit betrachtet wird, nutzt ein neuronales Netz zur Erzeugung von Vorschlägen für die Systemstellen bzw. Schlagwörter. Neuronale Netze sind sehr gut für die Generalisierung und Assoziation von Problemstellungen geeignet und kommen häufig im Bereich der Mustererkennung zur Anwendung [Kri07].

Ein Netz besteht aus mehreren Schichten, die jeweils eine bestimmte Anzahl an Neuronen haben [Ros58]. Diese Schichten sind in einer festen Reihenfolge angeordnet. Die erste Schicht ist die Eingabeschicht, die letzte Schicht die Ausgabeschicht und die mittleren Schichten heißen verdeckte Schichten. Die Neuronen einer Schicht sind vollständig mit den Neuronen der nächsten Schicht über Kanten verbunden. Zwischen den Neuronen einer einzelnen Schicht bestehen hingegen keinerlei Verbindungen. Eine Visualisierung eines neuronalen Netzes ist in Abbildung 4 zu sehen.

Jedes Neuron speichert einen numerischen Wert und die Kanten verfügen über Gewichte. Zur Berechnung mittels des neuronalen Netzes werden die Neuronen der Eingabeschicht zunächst von außen mit den Eingabewerten belegt. Danach werden schichtweise die Werte der Neuronen dieser Schicht berechnet, bis die Neuronen der Ausgabeschicht berechnet sind. Diese Werte bilden dann die Ausgabewerte des Netzes. Der Wert eines Neurons wird mittels der eingehenden Kanten bestimmt, dazu wird die Summe der Produkte der Kantengewichte mit den Werten der vorherigen Neuronen berechnet. Zum Schluss wird mit Hilfe einer Aktivierungsfunktion aus diese Summe der Wert des aktuellen Neurons berechnet. Bei dem hier vorgestellten Verfahren kommt eine Sigmoidfunktion zum Einsatz, um die Werte der Neuronen auf das Intervall $[0, 1]$ zu beschränken.

Die Kantengewichte werden vor der ersten Benutzung des Netzes mittels Trainingsda-

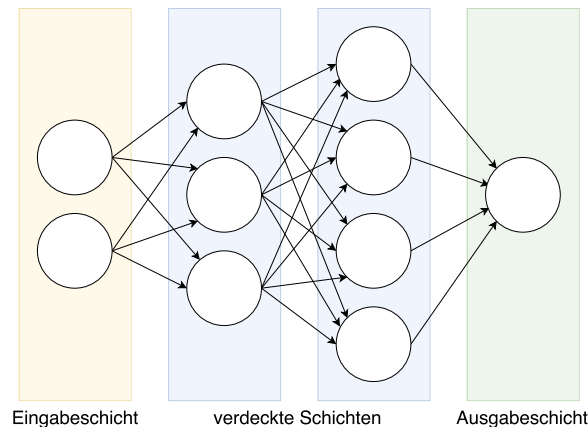


Abbildung 4: Beispiel eines neuronalen Netzes

ten trainiert. Dazu gibt es verschiedene Ansätze. Eine Möglichkeit stellt ein überwachtes Training dar, dabei werden sowohl die Werte der Eingabe- als auch der Ausgabeschicht vorgegeben. Die Kantengewichte werden dann so gewählt, dass ein möglichst kleiner Fehler entsteht. Auch in dieser Arbeit wird ein überwachtes Training in Form des Resilient Backpropagation Algorithmus [RB93] eingesetzt. Das Training wird beendet, sobald der Fehler der aktuellen Iteration im Verhältnis zum Fehler der vorigen Iteration weniger als 0,5% abweicht.

Zur Bestimmung von Systemstellen und Schlagwörtern ist wie bei dem KNN Klassifikator ein Feature-Vektor pro Buch notwendig. Die Werte des Feature-Vektors sind dabei die Eingabewerte der ersten Schicht des neuronalen Netzes. Für dieses Verfahren werden auch die auf Word2Vec basierenden Feature-Vektoren genutzt, welche in Abschnitt 3.2.1 erläutert werden. Somit wird jedes Buch durch ein 200-dimensionales Array beschrieben und die Eingabeschicht des Netzes hat folglich 200 Neuronen.

In der Ausgabeschicht des Netzes steht jedes Neuron für ein festes Schlagwort oder eine Systemstelle, sodass durch absteigende Sortierung der Werte der Ausgabeneuronen eine Vorschlagliste erzeugt werden kann. Dadurch bedingt ist es nicht möglich, mit einem neuronalen Netz gleichzeitig Systemstellen und Schlagwörter bestimmen zu können, sondern es muss für beide Fälle ein gesondertes Netz erstellt werden.

Das neuronale Netz wurde mittels des Accord.Net Frameworks umgesetzt [Sou14].

4 Bewertung der Verfahren

In diesem Kapitel werden die drei Verfahren aus Kapitel 3 näher betrachtet und deren Ergebnisse für verschiedene Testmengen ausgewertet. Dazu werden zunächst die Testdaten beschrieben und die Evaluationsmaße eingeführt. Danach wird jedes Verfahren detailliert analysiert und die Wahl für deren Parameter beschrieben.

4.1 Testdaten

Die Testdaten werden direkt aus JuLib bezogen. Insgesamt verfügt JuLib zum Stand 27.10.2016 über Daten von 127.121 gedruckten Büchern und 147.693 E-Büchern.

Um die Bestimmung der Systemstellen bewerten zu können, werden Datensätze mit Gold-Standard Werten für die Systemstellen benötigt. Mit diesen kann dann ein Vergleich mit den, durch die Verfahren, bestimmten Systemstellen durchgeführt werden. Von den gedruckten Büchern verfügen 121.048 über Systemstellen, hingegen nur 7.196 der E-Bücher. Bei den Schlagwörtern sieht es ähnlich aus. 65.364 gedruckte Bücher und 5.287 E-Bücher haben bereits Schlagwörter zugewiesen.

Generell muss festgehalten werden, dass sowohl die Annotation von Schlagwörtern sowie Systemstellen aller Einträge in JuLib jeweils durch eine Person erfolgt ist. Daher könnte es vorkommen, dass einzelne Daten fehlerhaft oder unvollständig sind und eine Abweichung von der Bestimmung durch die Verfahren nicht zwangsläufig einen Fehler der Verfahren darstellt. Für die weitere Bewertung wird aber die Richtigkeit der Werte in JuLib angenommen.

Zudem verfügen nicht alle Datensätze über alle Metadaten, so können beispielsweise

Buchart	Labels	Gesamtanzahl	Anzahl mit Inhaltsverzeichnis
Bücher	Schlagwörter	65.364	11.146
	Systemstellen	121.048	12.025
E-Bücher	Schlagwörter	5.287	1.652
	Systemstellen	7.196	1.671

Tabelle 4: Übersicht über die Testdaten Verteilung

se die Autoren oder das Inhaltsverzeichnis fehlen. Da die Verfahren auf die Verfügbarkeit eines digitalen Inhaltsverzeichnisses basieren, wird sich im Folgenden auf Bücher beschränkt, welche eines hinterlegt haben. In Tabelle 4 ist die Verteilung der Bücher dargestellt. Ersichtlich ist, dass nur ein kleiner Teil der Bücher über ein maschinenlesbares Inhaltsverzeichnis verfügen. Dabei ist aber zu beachten, dass fast alle Neuzugänge in JuLib mittlerweile ein Inhaltsverzeichnis haben, da bei den gedruckten Büchern dieses per OCR erfasst bzw. bei den E-Büchern mitgeliefert wird.

Label Kardinalität und Label Dichte

Zwei Maße, um die Verteilung der Label in den Gold-Standard Objekten zu untersuchen sind die Label Kardinalität und Label Dichte[TK07].

Buchart	Labels	LC	LD
Bücher	Schlagwörter	3,0172	0,000390
	Systemstellen	1,3851	0,000717
E-Bücher	Schlagwörter	3,4078	0,000441
	Systemstellen	1,5247	0,000790

Tabelle 5: Werte für $LC(D)$ und $LD(D)$ des Testdatensatzes

Die Label Kardinalität beschreibt die durchschnittliche Anzahl an Labeln Y_i pro Objekt x_i der Menge aller Objekte D . Im Kontext der Arbeit, also die durchschnittliche Anzahl an Schlagwörtern bzw. Systemstellen pro Buch. Mit Hilfe dieses Maßes lässt sich folglich eine Aussage über das Ausmaß der Multi-Label Nutzung treffen. Bei einer Label Kardinalität von 1 würde es sich um ein einfaches Single-Label System handeln. Umso höher der Wert für die Kardinalität ausfällt, desto komplexer ist die Zuweisung von Labeln zu den Objekten.

Die Label Kardinalität wird mit der folgenden Formel berechnet:

$$LC(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i| \quad (4.1)$$

Die Label Dichte nutzt auch die durchschnittliche Anzahl an Labeln pro Objekt, setzt diese allerdings noch in das Verhältnis zu der Anzahl aller Label L :

$$LD(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|} \quad (4.2)$$

Mit dieser Formel wird folglich die durchschnittliche Wahrscheinlichkeit berechnet,

mit der ein Buch eine bestimmte Systemstelle oder ein bestimmtes Schlagwort zugewiesen bekommt.

In der Tabelle 5 sind die Werte für die Schlagwörter und Systemstellen des Datensatzes dargestellt. Die Werte für die Label Kardinalität und Label Dichte wurden getrennt für gedruckte Bücher und E-Bücher berechnet, um Unterschiede zwischen den Medien zu erkennen. In beiden Fällen wurden nur Bücher einbezogen, welche einen Gold-Standard Wert für die Systemstellen bzw. Schlagwörter besitzen, also wenigstens einen Eintrag haben.

Die Werte für die Label Kardinalität und Label Dichte liegen im Vergleich zwischen gedruckten Büchern und elektronischen Büchern nahe beisammen. Trotzdem lässt sich feststellen, dass alle Werte bei den E-Büchern leicht höher ausfallen.

Durchschnittlich hat jedes gedruckte Buch drei Schlagwörter zugewiesen und 1,4 Systemstellen. Im Vergleich dazu hat ein E-Buch 3,4 Schlagwörter und 1,5 Systemstellen. Dabei ist bei allen Fällen die Label Dichte sehr niedrig. Ein Buch bzw. E-Buch hat nur zu ca. 0,04% ein bestimmtes Schlagwort und zu ca. 0,07% eine bestimmte Systemstelle.

4.2 Benutzte Evaluationsmaße

In diesem Abschnitt werden die verschiedenen Evaluationsmaße vorgestellt, mit denen die Verfahren aus Kapitel 3 bewertet und miteinander verglichen werden.

Die Menge an Testdaten wird im Folgenden als D bezeichnet. Diese besteht aus einzelnen Einträgen $(x_i, Y_i), i = 1 \dots |D|$, dabei ist x_i das eigentliche Buch und Y_i die Gold-Standard Menge an Labels. Ein Klassifikator H erzeugt daraus die Vorschlagslisten $Z_i = H(x_i)$.

Bei der Evaluation der Ergebnisse der Verfahren ist zu beachten, dass im Allgemeinen die Gold-Standard Werte für Schlagwörter und Systemstellen ohne Reihenfolge sind. Im Gegensatz dazu handelt es sich bei den Vorschlagslisten um nach Wahrscheinlichkeit absteigend sortierte Listen. Die beiden Maße Deckungsgrad und Treffsicherheit wurden im Rahmen dieser Arbeit speziell für den Anwendungsfall entwickelt und zugeschnitten und beachten genau diese Diskrepanz zwischen sortierter und nicht sortierter Liste. Bei den anderen Maßen handelt es sich um bereits bewährte Maße für Multi-Label Klassifikationen. In Tabelle 6 ist ein Beispiel für alle verwendeten Evaluationsmaße, mit Ausnahme der Treffsicherheit, aufgeführt.

Vorschlagsliste	Java (programming language), object oriented programming, world wide web, Internet programming languages, concrete, MS - Office, MS Visio, Internet
Gold-Standard	parallel algorithm, parallel processing, Java (programming language), network, Internet
Deckungsgrad	0,2
Hamming Loss	0,00116
Accuracy	0,18182
Precision	0,25
Recall	0,4

Tabelle 6: Beispiel für die Evaluationsmaße

4.2.1 Deckungsgrad

Bei dieser Bewertung wird die durchschnittliche Überlappung der Vorschlagsliste zum Gold-Standard betrachtet. Für die Berechnung des Deckungsgrades der Schlagwörter eines einzelnen Buches wird als erstes die absteigend sortierte Vorschlagsliste mit den relevantesten Schlagwörtern bzw. Systemstellen bestimmt.

$$\text{Coverage}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cup Z_{i,|Y_i|}|}{|Z_{i,|Y_i|}|} \quad (4.3)$$

Hat ein Buch n Gold-Standard Schlagwörter bzw. Systemstellen, so wird auch die Vorschlagsliste auf die ersten n Einträge beschränkt, im weiteren als $Z_{i,n}$ bezeichnet. Danach wird die Überlappung der beiden Listen berechnet und ins Verhältnis zu n gesetzt. Dies bedeutet, dass das prozentuale Vorkommen der Gold-Standard Werte in den ersten n Einträgen der Vorschlagsliste berechnet wird, im weiteren Deckungsgrad genannt. Die Deckungsgrade aller Bücher werden dann gemittelt und als Gesamtdurchschnitt ausgegeben. Damit ist der Deckungsgrad fast gleich der Precision (siehe 4.6), berücksichtigt aber, dass die Vorschlagsliste sortiert ist.

4.2.2 Hamming Loss

In [SS00] wird der Hamming Loss als Evaluationsmaß vorgestellt. Wie der Name bereits sagt, handelt es sich um eine Verlustfunktion. Dies bedeutet, dass ein Wert von 0 das optimale Ergebnis darstellt und bei steigendem Wert das Ergebnis schlechter wird. Der Hamming Loss stellt die symmetrische Differenz zwischen Gold-Standard und Vorschlagsliste ins Verhältnis zur Gesamtanzahl an Labels. Die Summe wird abschließend

normiert, damit das Maß Werte im Bereich $[0, 1]$ zurückgibt. Die Formel [TK07] für den Hamming Loss sieht folgendermaßen aus:

$$\text{HammingLoss}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|} \quad (4.4)$$

Hierbei bezeichnet L die Menge aller möglichen Labels. Der Operand Δ steht für die symmetrische Differenz zweier Mengen, welche vergleichbar mit dem *xor* in der booleschen Logik ist.

4.2.3 Accuracy, Precision und Recall

Die klassischen Maße Accuracy, Precision und Recall können auch bei der Multi-Label Klassifikation angewendet werden [GS04]. Die drei Maße vergleichen die Gold-Standard Vorgaben mit der Vorschlagsliste und stellen diese in verschiedene Beziehungen zueinander. Nach [TK07] werden im weiteren Verlauf der Arbeit folgende Definitionen genutzt:

$$\text{Accuracy}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \quad (4.5)$$

$$\text{Precision}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Z_i|} \quad (4.6)$$

$$\text{Recall}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i|} \quad (4.7)$$

Die Accuracy gibt Auskunft darüber, wie hoch der Anteil der richtigen Vorschläge an der Gesamtmenge der Vorschläge und Gold-Standard Werte ist. Das Maß bestraft also in erster Linie eine hohe Anzahl an vorgeblich falschen Vorschlägen, hängt aber auch von der Anzahl nicht gefundener Gold-Standard Werte ab. Bei einer Accuracy von 1 sind alle Vorschläge richtig und es sind alle Gold-Standard Werte enthalten.

Bei der Precision wird die Anzahl an richtigen Vorschläge durch die Anzahl aller Vorschläge geteilt. Sie berechnet somit den Anteil an richtigen Vorschlägen und lässt dabei die Anzahl an Gold-Standard Werten außer Acht. Bei einer Precision von 1 sind alle Vorschläge korrekt.

Zuletzt gibt der Recall den Anteil der gefundenen Gold-Standard Werte an. Falsche Vorschläge spielen für den Recall keine Rolle, so wurden bei einem Recall von 1 alle Gold-Standard Werte gefunden.

4.2.4 Treffsicherheit

Bei der Treffsicherheit wird nicht wie bei dem Deckungsgrad auf Basis der einzelnen Bücher bewertet, sondern auf Grundlage der einzelnen Label. Es wird pro Label berechnet, wie häufig das Label korrekterweise in der Vorschlagsliste vorkommt. Dabei findet, wie beim Deckungsgrad, eine Beschränkung der Vorschlagsliste auf so viele Stellen, wie das Buch an Gold-Standard Werte besitzt, statt.

Mit Hilfe der Treffsicherheit lassen sich einzelne Label identifizieren, für die die Verfahren weniger geeignet sind oder die Ergebnisse durch andere Faktoren beeinflusst werden.

Dieses Maß wird nicht zusammen mit den anderen Maßen betrachtet, sondern gesondert in einem eigenen Abschnitt, da die Treffsicherheit nur für die besten zuvor ermittelten Parameter der Verfahren durchgeführt wird.

4.2.5 Single-Label Betrachtung

Bei dieser Betrachtung werden nur diejenigen Bücher aus den Testdaten herangezogen, welche über eine einzige Systemstelle verfügen. Somit wird das Problem von Multi-Label auf Single-Label reduziert. Dieses wird allein für die Systemstellen ausgewertet, da durch den Anwendungsfall eine Reduktion auf ein einziges Schlagwort unrealistisch ist. Von den E-Büchern verfügen 4.412 und 86.327 der gedruckten Bücher über genau eine Systemstelle. Im Vergleich dazu haben nur 1.082 E-Bücher und 24.392 gedruckte Bücher ein einziges Schlagwort.

Da auch die Vorschläge der Verfahren auf Single-Label reduziert werden müssen, werden bei dieser Betrachtung nur noch jeweils der beste Vorschlag genommen und die anderen verworfen.

Als Evaluationsmaß kommen in der Single-Label Betrachtung die klassischen Maße Macro-Averaged Precision und Macro-Averaged Recall, welche den Durchschnitt über alle Label L berechnen, zur Anwendung [Yan99]:

$$\text{Precision}_{\text{macro}} = \frac{1}{|L|} \sum_{i=1}^{|L|} \frac{TP_i}{TP_i + FP_i} \quad (4.8)$$

$$\text{Recall}_{\text{macro}} = \frac{1}{|L|} \sum_{i=1}^{|L|} \frac{TP_i}{TP_i + FN_i} \quad (4.9)$$

TP_i ist die Anzahl der Vorkommen, wie oft das i -te Label richtig vorgeschlagen wurde, FP_i ist die Anzahl der falschen Vorschläge und FN_i die Anzahl, wie oft das Label i nicht vorgeschlagen wurde.

Die Macro-Averaged Precision ist der Anteil, wie oft ein Label richtigerweise vorgeschlagen wurde, an der Gesamtanzahl, wie oft das Label insgesamt vorgeschlagen wurde. Der Macro-Averaged Recall betrachtet die Vorschläge aus einem anderen Blickwinkel. Dieser gibt an, wie oft ein bestimmtes Label richtig vorgeschlagen wurde im Verhältnis zur Anzahl, wie oft es hätte vorgeschlagen werden müssen.

4.3 Auswertung für Systemstellen

In diesem Abschnitt werden die Vorschlagslisten für die Systemstellen der einzelnen Verfahren näher betrachtet und die Evaluationsmaße darauf angewendet. Aus den Testdaten werden dazu zunächst mehrere Testmengen erzeugt, sodass eine detaillierte Analyse ermöglicht wird. Die Testmengen sind in Tabelle 7 aufgelistet. Bei allen Datensätzen handelt es sich um diejenigen, die über Gold-Standard Werte für die Systemstellen verfügen.

Jede Testmenge wird in zehn gleich große Teile verteilt. Ein Teil dient für Abschnitt 4.5 als Testmenge, während die anderen neun hier genutzt werden, um eine Wahl der Parameter für die Verfahren zu ermöglichen. Für jede der Testmengen wird mittels der neun Teile eine Kreuzvalidierung durchgeführt. Für jede Kreuzvalidierung wird in den folgenden Abschnitten jeweils der Durchschnitt der neun Durchläufe angegeben.

	Beschreibung	Datensätze	Verschiedene Label
Testmenge A1	Gedruckte Bücher mit Inhaltsverzeichnis	12.025	1.555
Testmenge A2	E-Bücher mit Inhaltsverzeichnis	1.671	827
Testmenge A3	Bücher mit Inhaltsverzeichnis	13.696	1.583
Testmenge A4	Testmenge 3 mit reduzierter Systematik	13.696	16

Tabelle 7: Testmengen für Systemstellen

Die Testmenge A4 stellt einen Sonderfall dar. In Testmenge A4 werden alle Gold-Standard Systemstellen sowie die Vorschläge auf die erste Stelle gekürzt und dann miteinander verglichen und ausgewertet. Dies verringert die Anzahl der unterschiedlichen Label und vereinfacht die Problemstellung.

Die Einzelergebnisse der Auswertungen sind in Anhang C einzusehen.

4.3.1 BookGraph

Bei dem BookGraph Verfahren gibt es zwei Parameter, welche für ein besseres Ergebnis verändert werden können. Dies ist zum einen der Schwellwert, ab welchem ein Pfad im Graph akzeptiert wird und zum anderen ein boolescher Wert, der angibt, ob ein dynamischer Schwellwert verwendet wird. Der dynamische Schwellwert senkt, bei weniger

als zehn Vorschlägen, den Schwellwert graduell um 0,01, bis mindestens zehn Vorschläge gefunden werden oder der Schwellwert 0 erreicht. Der dynamische Schwellwert wird dabei nach jedem Buch zurückgesetzt.

Generell gilt, dass bei niedrigerem Schwellwert, mehr Vorschläge gemacht werden. Allerdings folgt aus einem sehr kleinen Schwellwert nicht, dass dadurch alle in den Trainingsdaten vertretenen Label vorgeschlagen werden. Dies liegt zum einen daran, dass bei der Bestimmung der Systemstellen nur die direkten Nachbarn der Wörter des Inhaltsverzeichnisses betrachtet werden und zum anderen daran, dass der Graph nicht zwingenderweise zusammenhängend ist, die Testmenge A3 besteht beispielsweise aus zwölf Teilgraphen, und die Bestimmung nur auf einem kleinen Teilgraphen erfolgen könnte.

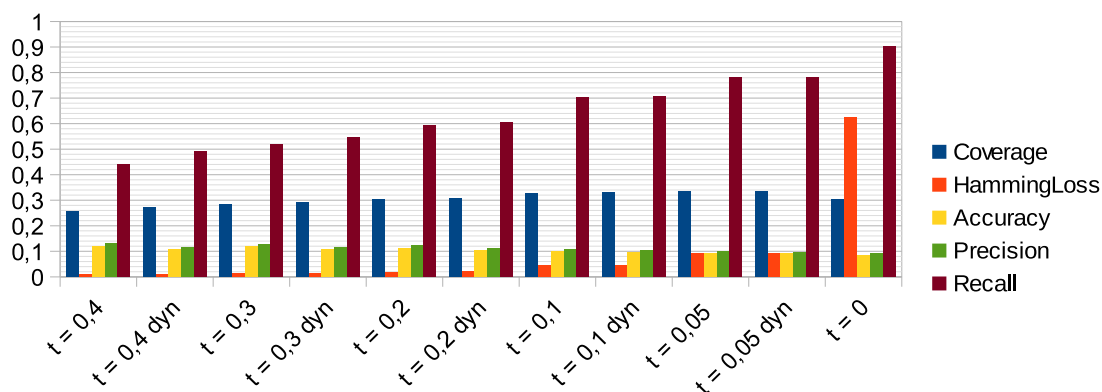


Abbildung 5: Ergebnisse Systemstellen BookGraph Testmenge A1

Zunächst wird die Testmenge A1 betrachtet. In Grafik 5 sind die Ergebnisse für die verschiedenen Ähnlichkeitsmaße aufgeführt. Zu sehen sind die Ergebnisse für die verschiedenen Schwellwerte, jeweils mit dynamischer Schwelle aktiviert und deaktiviert. Am auffälligsten ist der steigende Recall, von 0,4408 auf 0,9005, bei fallendem Schwellwert. Dies liegt vor allem daran, dass dadurch die Vorschlagsliste größer wird und damit die Wahrscheinlichkeit steigt, dass die Gold-Standard Werte enthalten sind.

Auch der HammingLoss steigt aufgrund der größeren Vorschlagsliste an. Die Anzahl richtiger Vorschläge steigt nicht im Verhältnis zu der Gesamtanzahl an Vorschlägen und somit wird die symmetrische Differenz größer, während die Labelanzahl eine Konstante ist. Der HammingLoss steigt dadurch von 0,0081 auf 0,6231 an.

Accuracy und Precision fallen bei sinkendem Schwellwert leicht von 0,1202 auf 0,0848 bzw. von 0,1318 auf 0,0911. Hier ist der Effekt nicht so deutlich, da das Verhältnis von richtigen Vorschlägen zu falschen Vorschlägen recht stabil ist.

Der Deckungsgrad (Coverage) ist das interessanteste Maß. Dieser gibt an, wie gut die

ersten paar Vorschläge sind (siehe Formel 4.3) und ist damit für den späteren Praxiseinsatz am relevantesten. Die Sachbearbeiter der ZB möchten sich nicht durch eine lange Liste an Vorschlägen arbeiten, sondern in den obersten Einträgen passende Vorschläge vorfinden (weiteres siehe Abschnitt 4.8).

Der Coverage steigt kontinuierlich bei fallendem Schwellwert von 0,2576 auf 0,3342 für $t = 0,05$ an. Nur ein Schwellwert von 0 lässt den Deckungsgrad wieder auf 0,303 sinken.

Bei allen Werten für den Schwellwert ist zu sehen, dass eine dynamische Schwelle bessere Werte für Coverage und Recall liefert als ein fester Schwellwert. Somit sollte bei der Testmenge A1 ein möglichst kleiner dynamischer Schwellwert gewählt werden, um die besten Vorschläge zu liefern.

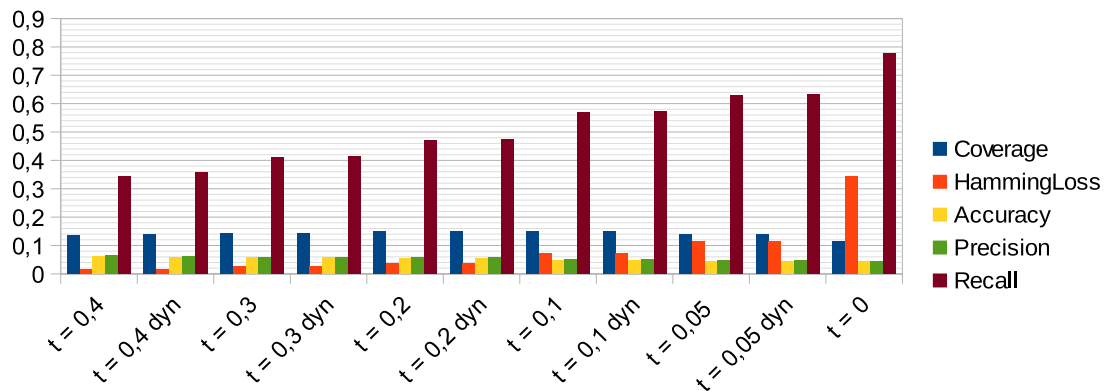


Abbildung 6: Ergebnisse Systemstellen BookGraph Testmenge A2

Für die Testmenge A2 (siehe Abbildung 6) ergibt sich ein ähnliches Bild. Aufgrund des geringeren Umfangs an Datensätzen in der Testmenge A2 fallen die Ergebnisse entsprechend niedriger aus, da das Verfahren dadurch über weniger Trainingsdaten verfügt. Einzig der Coverage verhält sich im Gegensatz zur Testmenge A1 anders. Dieser ist nicht kontinuierlich steigend bei fallendem Schwellwert, sondern hat sein Maximum von 0,1505 bei einer dynamischen Schwelle von 0,2. Auch hierbei sind die dynamischen Schwellwerte besser als ihre festen Gegenstücke.

Die Testmenge A3 als Kombination aus den vorigen beiden Testmengen hat die meisten Testdaten und es werden dadurch auch bessere Ergebnisse erzielt. Bei dieser Menge zeigen die Maße ein leicht anderes Verhalten (Abbildung 7). Während Coverage, Recall bei fallendem Schwellwert kontinuierlich steigen, trifft dies für Precision und Accuracy nicht zu. Beide Maße haben ihr Maximum bei einem festen Schwellwert von 0,3, dabei liegt die Precision bei 0,1476 und die Accuracy bei 0,136.

Insgesamt lässt sich für die dritte Testmenge aber auch zusammenfassen, dass eine

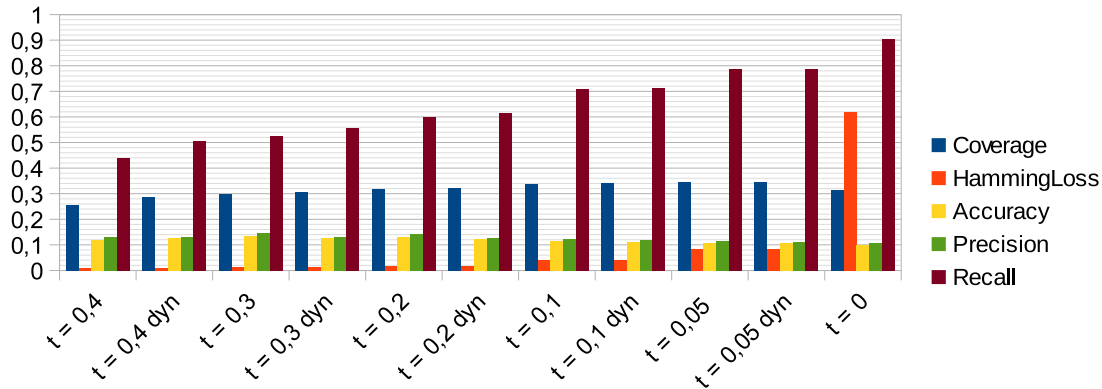


Abbildung 7: Ergebnisse Systemstellen BookGraph Testmenge A3

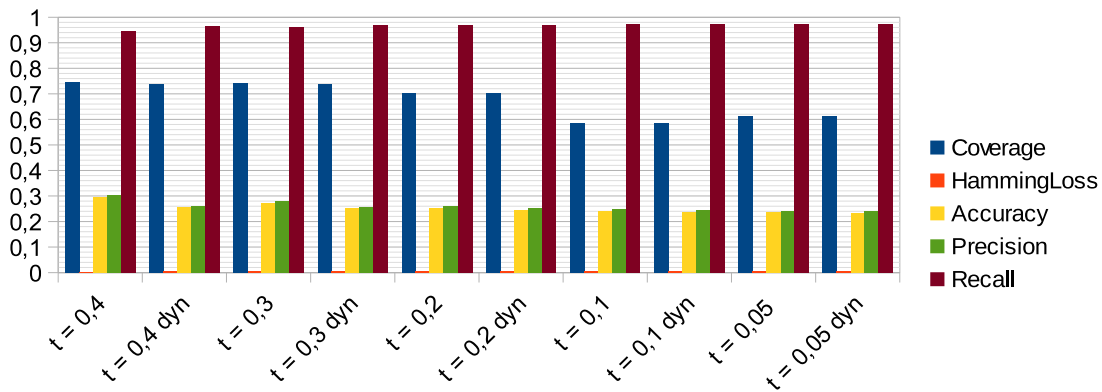


Abbildung 8: Ergebnisse Systemstellen BookGraph Testmenge A4

dynamische Schwelle von 0,05 die besten Gesamtergebnisse liefert.

Bei der Betrachtung einer reduzierten Systematik in Testmenge A4 fallen die Ergebnisse für BookGraph deutlich anders aus als bei den vorigen Testmengen. In Abbildung 8 sind die Ergebnisse dargestellt. Wurden bei den vorigen Testmengen noch schlechtere Ergebnisse bei zu hohem Schwellwert erzielt, so ist hier der beste Wert für den Deckungsgrad bei einem Schwellwert von 0,4 und einem Wert von 0,7473. Der Recall verhält sich wie erwartet, liegt aber bei allen Schwellwerten bei einem sehr hohen Niveau zwischen 0,9466 und 0,9733. Auch die übrigen Maße verschlechtern sich bei niedrigerem Schwellwert. Insgesamt fallen die Ergebnisse für die Testmenge A4, bedingt durch die kleinere Anzahl möglicher Label, wesentlich besser aus als bei den anderen Testmengen.

4.3.2 k -Nächste-Nachbarn

Der k -Nächste-Nachbarn Klassifikator besitzt, wie in Abschnitt 3.2 beschrieben, nur den Parameter k , welcher die Anzahl nächster Nachbarn für die Auswertung angibt. Bei steigendem Wert für k werden somit mehr Vorschläge gemacht, bis die Menge an Vorschlägen alle Labels umfasst, die in den Trainingsdaten vorkommen.

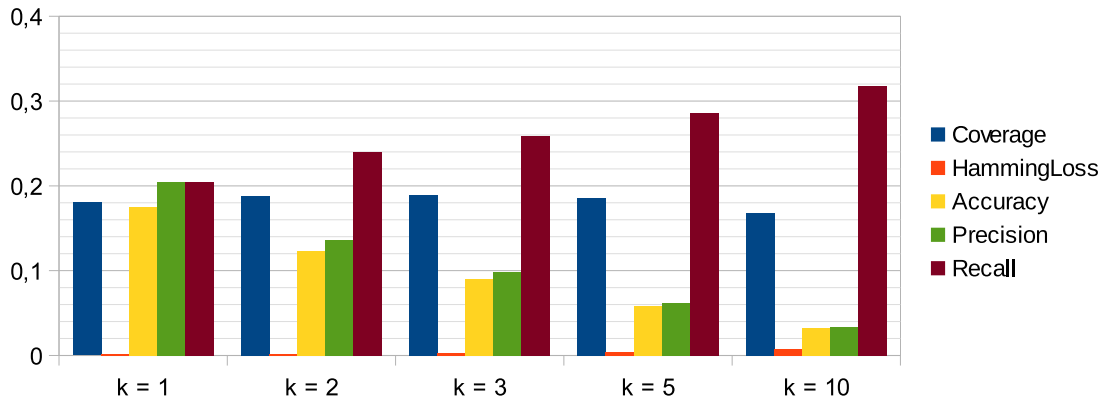


Abbildung 9: Ergebnisse Systemstellen KNN Testmenge A1

Die Ergebnisse für die erste Testmenge sind in Abbildung 9 dargestellt. Wie zu erwarten fällt der Recall bei $k = 10$ mit 0,3177 am höchsten aus und fällt bis zu $k = 1$ auf 0,2041 ab. Dahingegen steigen bei einer kleineren Nachbarschaft die Precision und Accuracy. Bei einem $k = 1$ beträgt die Accuracy 0,1748 und die Precision 0,2048. Auch der HammingLoss verbessert sich bei kleinerem k .

Der Deckungsgrad hat sein Maximum bei $k = 3$ mit einem Wert von 0,1892. Ein kleineres k schränkt die Nachbarn zu stark ein und einzelne Ausreißer verschlechtern das Ergebnis, wohingegen ein größeres k für eine zu starke Glättung sorgt.

Ein ähnliches Bild ergibt sich bei der Betrachtung der Testmenge A2 (Abbildung 10). Auch hier fällt wie beim BookGraph Verfahren auf, dass im Vergleich zu der Testmenge A1 deutlich schlechtere Ergebnisse erzielt werden. Dies liegt wiederum an der geringen Anzahl an Datensätzen in der zweiten Menge, so dass nur wenige Daten für ein Training zur Verfügung stehen. Das Maximum des Coverage ist 0,9411 bei einem Umkreis von 2 Nachbarn. Der Recall steigt erwartungsgemäß für eine größeres k aufgrund der größeren Vorschlagsliste und hat bei $k = 10$ einen Wert von 0,17772. Precision, Recall und HammingLoss verbessern sich deutlich bei kleinerem k .

Bei der Testmenge A3, als Kombination der ersten beiden Testmengen, verhalten sich alle Werte gleich. Da bei dieser Menge die meisten Datensätze zur Verfügung stehen, fal-

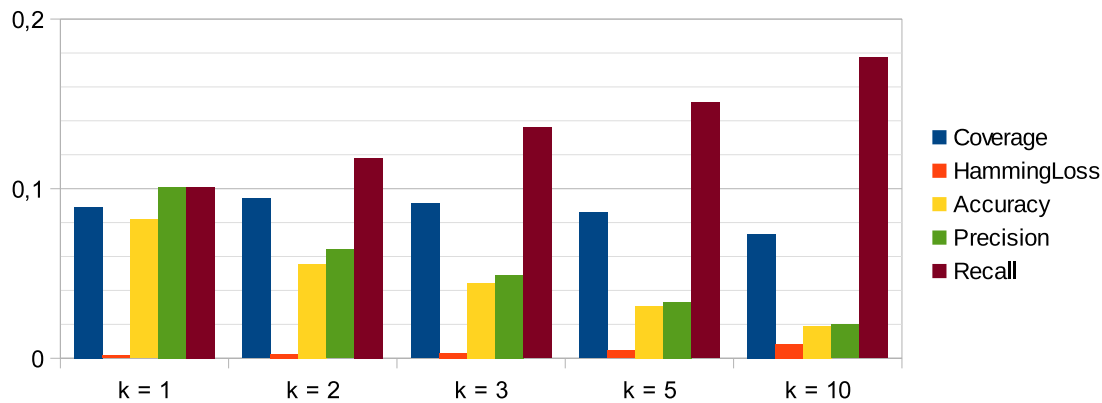


Abbildung 10: Ergebnisse Systemstellen KNN Testmenge A2

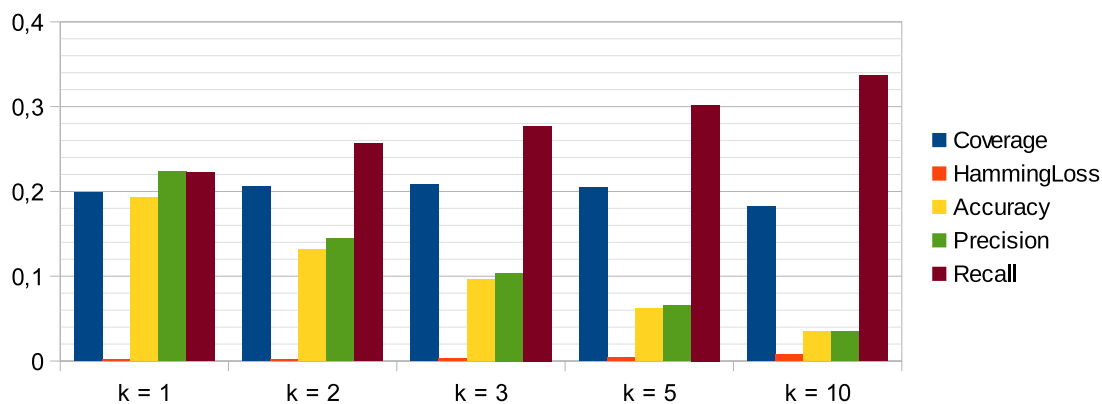


Abbildung 11: Ergebnisse Systemstellen KNN Testmenge A3

len die Ergebnisse besser aus (siehe Abbildung 11). Der Coverage erreicht sein Maximum von 0,2081 bei $k = 3$. Die Precision und Accuracy steigen von 0,0353 bzw. 0,0343 bei einer Nachbarschaft von 10 bis auf 0,2231 bzw. 0,1933 bei $k = 1$ an. Damit liegt für $k = 1$ die Precision leicht oberhalb des Recall, welcher 0,2219 beträgt.

Somit sollten für ein möglichst gutes Ergebnis des k -Nächste-Nachbarn Verfahrens $k = 3$ gewählt werden.

Zuletzt wird auch für dieses Verfahren die Testmenge A4 mit der reduzierten Systematik betrachtet (siehe Abbildung 12), um das Verhalten des Klassifikators bei einer kleinen Anzahl möglicher Labels zu testen.

Alle Werte verbessern sich im Vergleich zur vollen Systematik in Testmenge A3. So ist der beste Wert für den Deckungsgrad 0,6139 bei $k = 2$. Wie zuvor haben Precision und Recall ihre Maxima beim kleinsten Wert für k und der HammingLoss hat dort sein

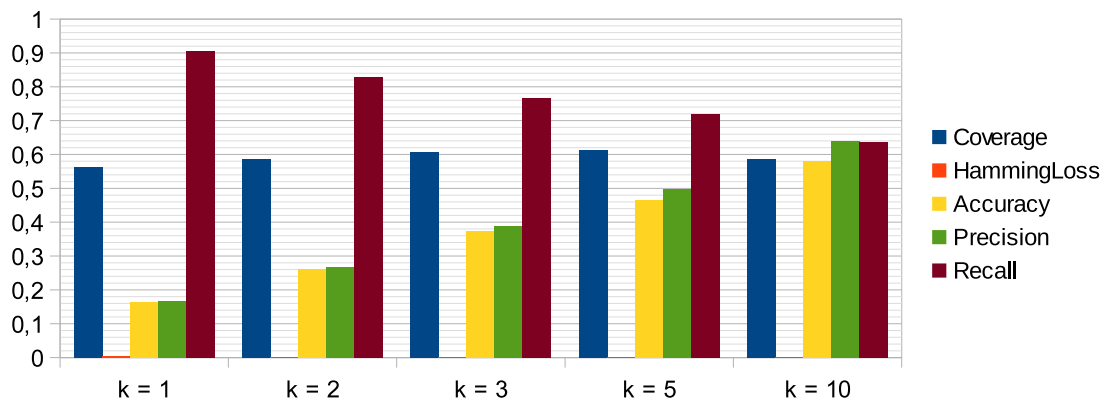


Abbildung 12: Ergebnisse Systemstellen KNN Testmenge A4

Minimum von 0,0005.

Der Recall hat sein Maximum bei 0,9044, was an der sehr kleinen Menge möglicher Labels liegt. Werden, wie bei $k = 10$, viele Bücher berücksichtigt, so ist die Wahrscheinlichkeit groß, alle Labels in der Vorschlagsliste zu haben.

4.3.3 Neuronales Netz

Die Auswertung für das neuronale Netz ist schwieriger zu evaluieren. Das neuronale Netz besitzt mehrere voneinander abhängige Parameter. Zum einen muss die Anzahl an verdeckten Schichten und zum anderen für jede der Schichten die Anzahl an Neuronen angegeben werden. Dadurch ergibt sich ein sehr großer Freiheitsgrad. Zusätzlich ist das Training des Netzes nicht deterministisch, sondern hat zufällige Bestandteile. So kann es vorkommen, dass das Training fehlschlägt, indem die Optimierungsfunktion in eine unpassende lokale Extremstelle läuft und das neuronale Netz danach eine schlechte Performance aufweist. Die hier vorgestellten Konfigurationen der verdeckten Schichten stellen nur eine Auswahl zur besseren Übersichtlichkeit dar, darüber hinaus wurden im Rahmen der Arbeit weitere getestet.

Beim neuronalen Netz sind grundsätzlich alle Systemstellen in der Ausgabeschicht vertreten und werden mit einem Wert versehen. Somit ist die Länge der Vorschlagsliste immer maximal lang und alle Maße außer des Deckungsgrades sind für alle Konfigurationen gleich und ohne Aussagekraft.

Die Ergebnisse für die Testmenge A1 sind in Abbildung 13 dargestellt. Auf der x-Achse sind die verschiedenen Konfigurationen für die verdeckten Schichten aufgetragen. Die Zahlen geben an, wie viele Neuronen in der verdeckten Schicht enthalten sind. Kommen

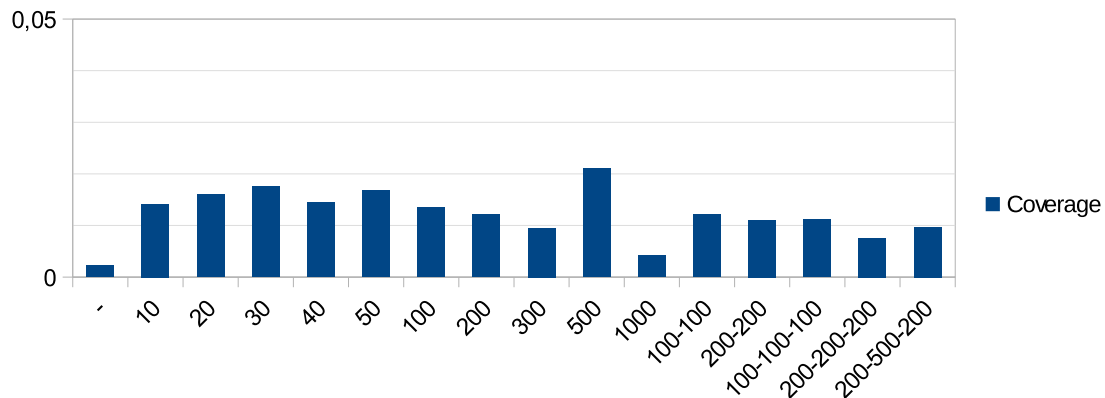


Abbildung 13: Ergebnisse Systemstellen Netz Testmenge A1

mehr als eine verdeckte Schicht zur Anwendung, so sind die Neuronenanzahlen mit einem „-“ voneinander getrennt.

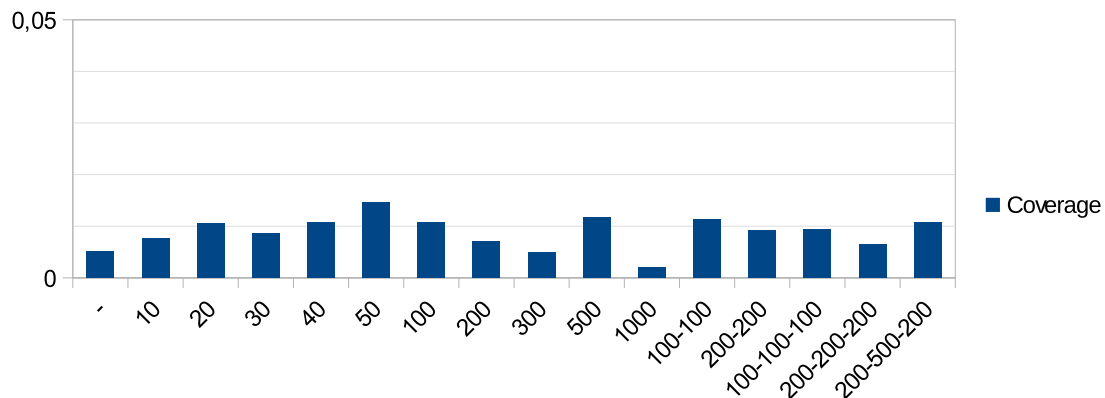


Abbildung 14: Ergebnisse Systemstellen Netz Testmenge A2

Deutlich zu erkennen sind die durchweg schlechten Ergebnisse für den Deckungsgrad. Das Maximum von 0,021 wird für die Konfiguration mit 500 Neuronen in einer einzigen verdeckten Schicht erreicht. Bei einer erneuten Berechnung der Werte könnte es allerdings, aufgrund der zufälligen Komponenten beim Training des Netzes, dazu kommen, dass eine andere Konfiguration ein besseres Ergebnis liefert. Von den hier dargestellten Werten lässt sich aber eine Tendenz herleiten. Eine Konfiguration mit nur einer verdeckten Schicht und wenigen Neuronen in dieser, liefert bessere Ergebnisse als mehrere Schichten oder eine hohe Anzahl an Neuronen. Das Maximum bei 500 Neuronen stellt bei der Betrachtung der umliegenden Werte einen einzelnen Ausreißer dar.

Bei der Betrachtung der Testmenge A2 (Abbildung 14) ist dieser Ausreißer auch zu

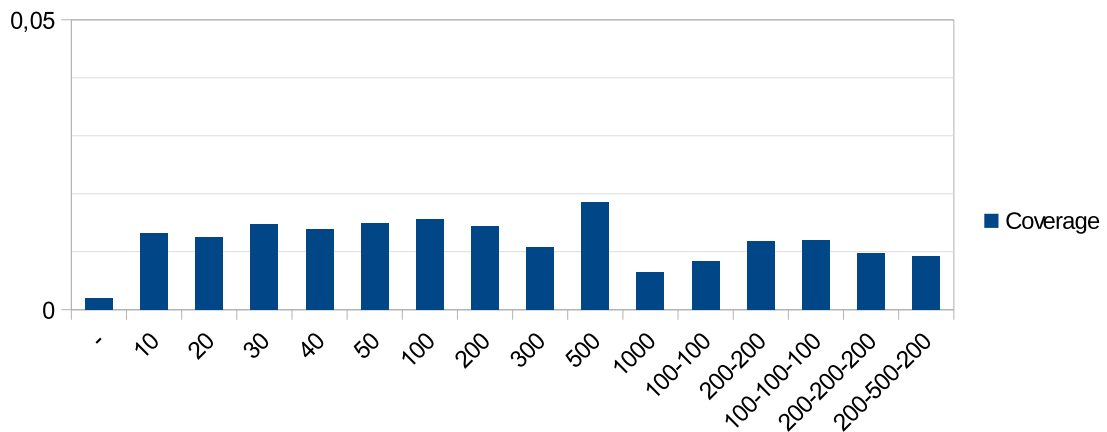


Abbildung 15: Ergebnisse Systemstellen Netz Testmenge A3

erkennen, wenn auch nicht so deutlich ausgeprägt. Generell fallen alle Ergebnisse in dieser Testmenge niedriger aus und die Werte liegen näher beisammen. Der maximale Deckungsgrad von 0,0147 wird für 50 Neuronen in einer verdeckten Schicht erreicht.

Auch bei der Testmenge A3, dargestellt in Abbildung 15, zeigen sich, im Vergleich zu den beiden vorigen Klassifikatoren, Auffälligkeiten. Die erzielten Ergebnisse der Testmenge A3 liegen unterhalb denen der Testmenge A1. Dies ist ungewöhnlich, da eigentlich bei steigender Anzahl an Datensätzen für das Training die Klassifikationsergebnisse besser werden sollten. Dies könnte zum einen durch Überanpassung, aber zum anderen auch durch zu starke Diversität innerhalb der Trainingsdaten herrühren. Das Maximum liegt wie bei der Menge A1 bei 500 Neuronen, erreicht aber nur einen Wert von 0,0185.

Die sehr hohe Anzahl an möglichen Labels und damit einer hohen Anzahl an Neuronen in der Ausgabeschicht kann vom Netz nicht effektiv gelernt werden. Dies zeigt sich vor allem bei der Betrachtung der vierten Testmenge A4.

Für die reduzierte Systematik und der dadurch reduzierten Größe der Ausgabeschicht kann das neuronale Netz wesentlich bessere Ergebnisse liefern (siehe Abbildung 16). So liegt das Maximum des Coverage bei 0,395 für 300 Neuronen.

4.4 Auswertung für Schlagwörter

Die Auswertung für die Schlagwortvorschläge erfolgt analog zur der Auswertung für die Systemstellen. Auch hier werden die Testmengen in zehn Teile geteilt, wobei ein Teil für den Test dient. Mit den übrigen wird eine Kreuzvalidierung durchgeführt. Die in diesem Abschnitt genutzten Testmengen sind in Tabelle 8 aufgelistet.

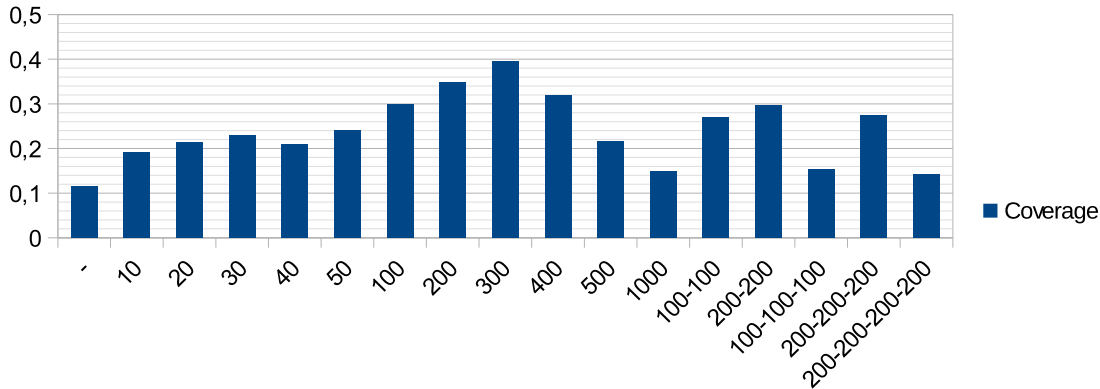


Abbildung 16: Ergebnisse Systemstellen Netz Testmenge A4

	Beschreibung	Datensätze	Verschiedene Label
Testmenge B1	Gedruckte Bücher mit Inhaltsverzeichnis	11.146	5.130
Testmenge B2	E-Bücher mit Inhaltsverzeichnis	1.652	2.312
Testmenge B3	Bücher mit Inhaltsverzeichnis	12.798	5.292

Tabelle 8: Testmengen für Schlagwörter

Generell ist zu beachten, dass die Menge an Schlagwörtern, wie in Abschnitt 4.1 dargestellt, deutlich umfangreicher ist als die Menge an Systemstellen. Aufgrund der größeren Auswahl an Möglichkeiten sind somit die erwarteten Ergebnisse der Verfahren tendenziell schlechter als bei der Auswertung für die Systemstellen.

Auch für die Auswertung der Schlagwörter sind die Einzelergebnisse im Anhang zu finden (Anhang D).

4.4.1 BookGraph

Das BookGraph Verfahren arbeitet bei der Bestimmung von Schlagwörtern geringfügig anders als bei der Bestimmung der Systemstellen. Die Parameter des Verfahrens bleiben allerdings gleich. Auch hier ist ein Schwellwert t vorhanden, welcher optional dynamisch gesenkt wird, sollten weniger als zehn Vorschläge gemacht werden. Es gelten im Weiteren die gleichen Anmerkungen zum Parameter t wie in Abschnitt 4.3.2.

Die Evaluationsmaße zeigen bei den Schlagwörtern die gleichen Verhalten wie zuvor bei den Systemstellen. Bei kleinerem Schwellwert sinken die Precision und die Accuracy, während der Deckungsgrad und der Recall steigen.

Die Werte für die Testmenge B1 sind in Abbildung 17 dargestellt. Bei allen Schwellwerten ist zu sehen, dass die dynamische Absenkung die Werte verbessert. Besonders

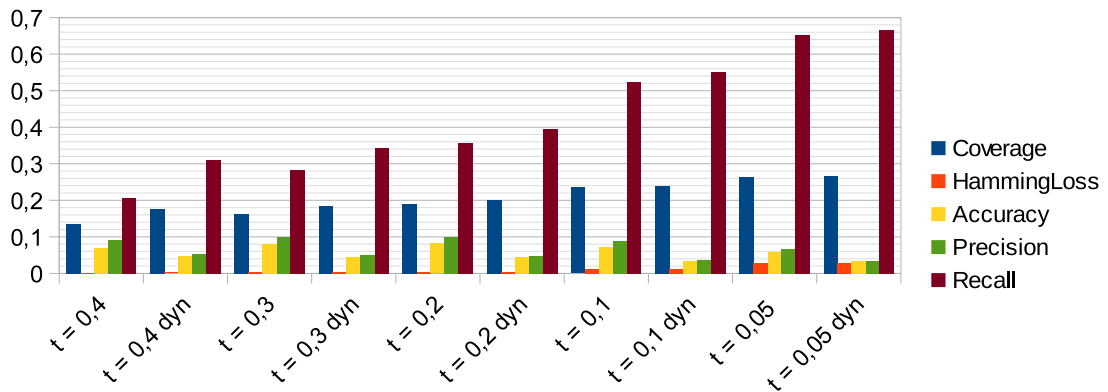


Abbildung 17: Ergebnisse Schlagwörter BookGraph Testmenge B1

deutlich wird dies bei den Schwellwerten 0,3 und 0,4. Dort fällt bei einer fixen Schwelle die Leistung des Verfahrens immens ab, da die hohe Schwelle zu wenig Kanten im Graphen zulässt und somit die Vorschlagsliste sehr eingeschränkt wird.

Der höchste Deckungsgrad wird so auch für $t = 0,05$ und einem Wert von 0,2658 erreicht. Bei diesem Schwellwert ist auch der Recall mit 0,666 am höchsten. Accuracy und Precision haben hingegen ihre Maxima bei $t = 0,2$, da dieser Schwellwert die beste Balance zwischen einer zu umfangreichen Vorschlagsliste und einer zu starken Einschränkung darstellt.

Bei der Testmenge B2 (Abbildung 18) sind die Auffälligkeiten bei hohen, nicht dynamischen Schwellwerten nicht vorhanden. Die Werte für Deckungsgrad und Recall verbessern sich kontinuierlich bei fallendem Schwellwert und betragen maximal 0,1388 bzw. 0,541. Wie bei Testmenge B1, sind bei einem festen Schwellwert von 0,2 Accuracy und Precision am höchsten. Der HammingLoss ist minimal bei dem größten hier aufgeführten Schwellwert von 0,4 und beträgt dort 0,0032.

Die Testmenge B3, als Kombinationen der Mengen B1 und B2, zeigt wiederum die Merkmale wie die ersten Testmengen. Die festen Schwellwerte 0,3 und 0,4 führen zu deutlich schlechteren Werten als ihre dynamischen Gegenstücke. Für die dritte Testmenge kann für den Deckungsgrad 0,2823 und für den Recall 0,6758 erzielt werden. Beides für einen dynamischen Schwellwert von 0,05. Abermals fallen für $t = 0,02$ die Accuracy und Precision am besten aus, bei kleineren Schwellwerten verschlechtern die umfangreichen Vorschlagslisten die Ergebnisse zusehends.

Zusammenfassend sollte zur Bestimmung der Schlagwörter für das BookGraph Verfahren ein dynamischer Schwellwert von 0,05 gewählt werden, um die besten Vorschläge zu bekommen.

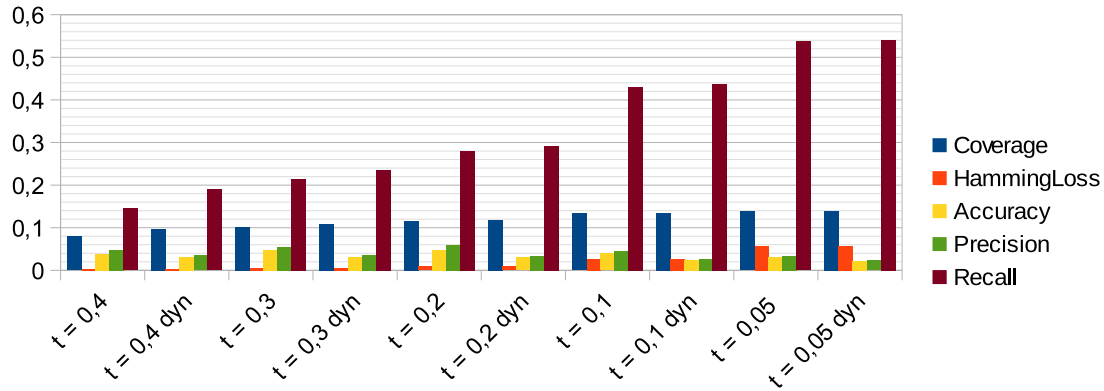


Abbildung 18: Ergebnisse Schlagwörter BookGraph Testmenge B2

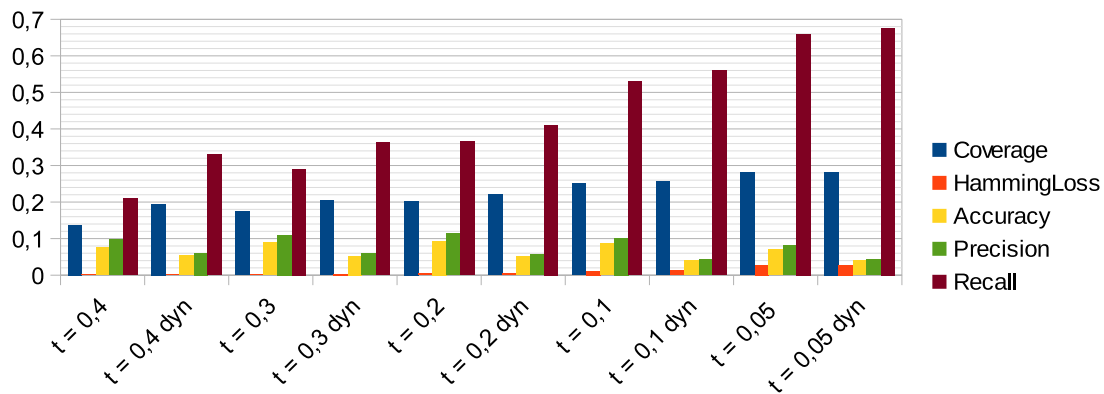


Abbildung 19: Ergebnisse Schlagwörter BookGraph Testmenge B3

4.4.2 k -Nächste-Nachbarn

Für die Bestimmung von Schlagwörtern mittels des k -Nächste-Nachbarn Klassifikators gelten die gleichen Anmerkungen zum Parameter k wie bei der Bestimmung der Systemstellen. Bei steigendem Wert von k werden mehr Nachbarn betrachtet und folglich mehr Vorschläge für mögliche Schlagwörter gemacht.

In Abbildung 20 sind die Ergebnisse der Auswertung für die Testmenge B1 dargestellt. Das Verhalten des Verfahrens bei der Bestimmung der Schlagwörter ist sehr ähnlich zu der Bestimmung der Systemstellen (vgl. 4.3.2). Der Deckungsgrad hat sein Maximum von 0,1402 bei $k = 2$. Bei einem kleineren oder größeren Wert für k verschlechtert sich wieder der Deckungsgrad, da einzelne Ausreißer bzw. eine Glättung über zu verschiedene Bücher auftritt. Die Precision verbessert sich von 0,0238 auf 0,1636 bei fallendem k . Auch die Accuracy steigt von 0,0224 auf 0,121. Gleichzeitig verschlechtert sich der Recall

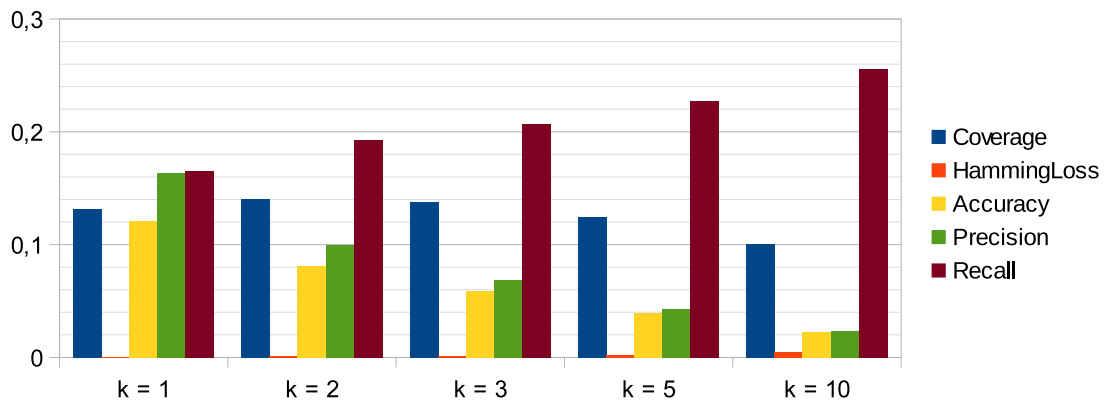


Abbildung 20: Ergebnisse Schlagwörter KNN Testmenge B1

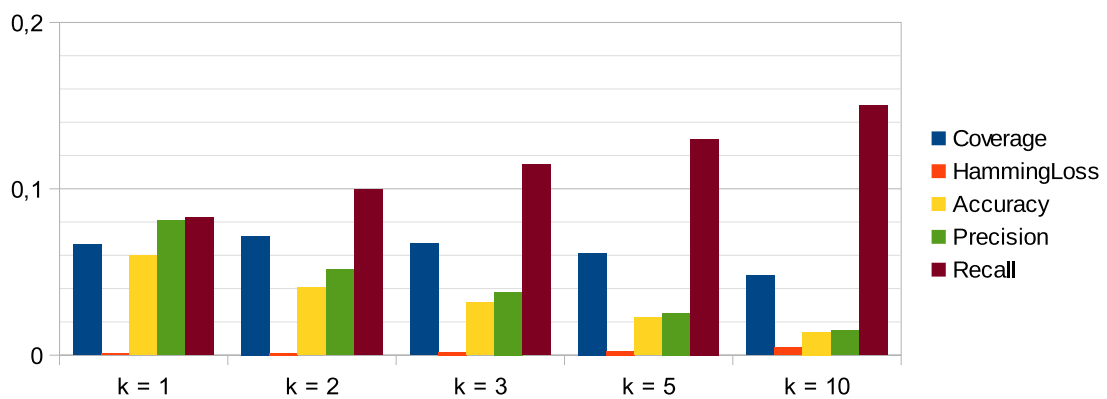


Abbildung 21: Ergebnisse Schlagwörter KNN Testmenge B2

aufgrund des geringeren Umfangs an Vorschlägen von 0,2554 auf 0,1651. Dies führt aber auch zu einer Verbesserung des HammingLoss bei kleinen Werten für k . Sein Minimum beträgt 0,0008.

Für die Auswertung auf Grundlage der E-Bücher in Testmenge B2 (Abbildung 21) gelten die gleichen Regelmäßigkeiten wie für die Menge B1. Generell fallen hier die Ergebnisse wieder schlechter aus, da die Menge B2 deutlich kleiner ist und somit weniger Datensätze für ein Training des Klassifikators vorhanden sind.

Der Deckungsgrad erreicht sein Maximum wieder für $k = 2$ bei einem Wert von 0,0715. Der Recall, mit einem Maximum von 0,1502, fällt auch sehr niedrig aus. Dies bedeutet, dass bei den E-Büchern auch bei größeren Vorschlagslisten, selten die Gold-Standard Schlagwörter enthalten sind. Die niedrigen Ergebnisse setzen sich bei Accuracy und Precision fort, deren größten Werte 0,06 und 0,0809 betragen. Im Gegensatz dazu zeigt der

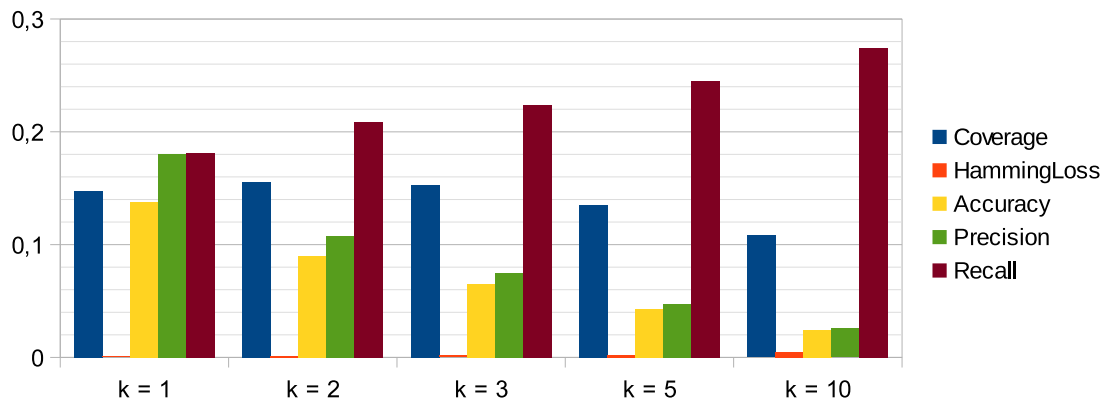


Abbildung 22: Ergebnisse Schlagwörter KNN Testmenge B3

HammingLoss keine Auffälligkeiten im Vergleich zur Testmenge B1 und liegt auf ähnlichem Niveau.

Bei der dritten Testmenge werden erwartungsgemäß die besten Ergebnisse erzielt, da sie die meisten Datensätze für das Training bereitstellt. Der höchste Wert für den Deckungsgrad ist 0,1555 und wird, wie bei den ersten beiden Mengen, bei $k = 2$ erreicht. Der Recall beträgt maximal 0,2743. Das heißt, dass selbst bei der größten Testmenge nicht einmal ein Drittel der Gold-Standard Schlagwörter in den Vorschlägen enthalten und nur ca. jedes sechste Schlagwort in den vorderen Vorschlägen vertreten ist. Für $k = 1$ erreichen Precision und Accuracy zwar 0,18 bzw. 0,1375 und der HammingLoss ist sehr niedrig mit einem Wert von 0,000796, dies bedeutet aber nur, dass das Verhältnis von richtigen Vorschlägen zu allen Vorschlägen akzeptabel ist.

4.4.3 Neuronales Netz

Auch für diese Auswertung wurden für das neuronale Netz verschiedene Konfigurationen für die Anzahl Schichten und deren Neuronen ausprobiert. In den folgenden Abbildungen ist wieder eine Auswahl der getesteten Konfigurationen dargestellt, sowie nur die Werte für den Deckungsgrad, da die anderen Maße keine Aussagekraft besitzen (siehe Abschnitt 4.3.3).

Bereits bei der Auswertung für die Systemstellen war zu erkennen, dass das neuronale Netz Probleme mit einer großen Anzahl an Labels hat. So funktionierte die Klassifikation für die reduzierte Systematik auch deutlich besser. Allerdings ist die Menge an Schlagwörtern nochmals deutlich größer als die der Systemstellen und folglich ist bei dieser Auswertung mit schlechteren Ergebnissen zu rechnen.

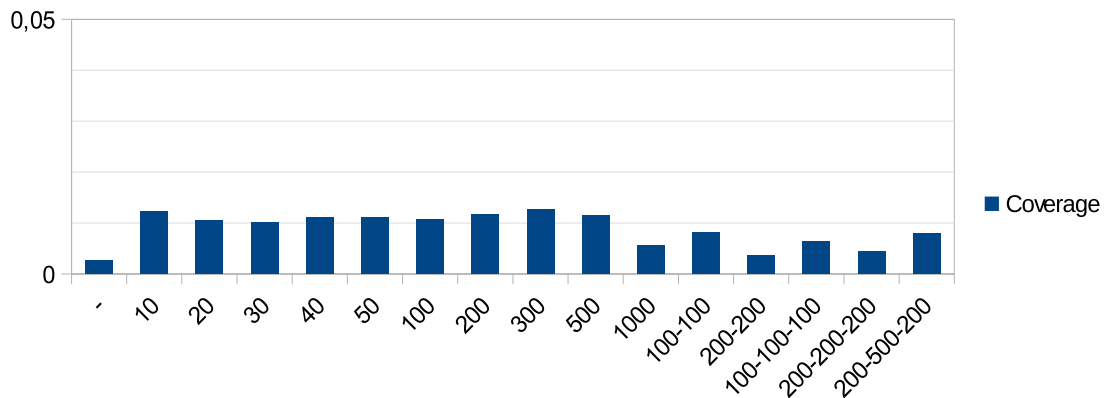


Abbildung 23: Ergebnisse Schlagwörter Netz Testmenge B1

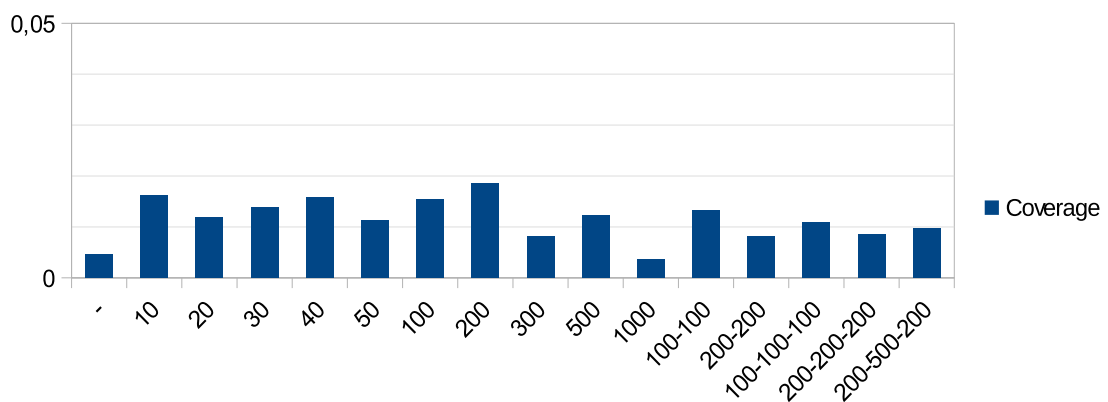


Abbildung 24: Ergebnisse Schlagwörter Netz Testmenge B2

Als erstes wird wieder die Testmenge B1 betrachtet, die Ergebnisse sind in Abbildung 23 dargestellt. Auf den ersten Blick zeigen sich keine Spitzen bei den Ergebnissen, sodass sich keine Konfiguration deutlich abheben kann. Die Konfigurationen mit einer einzigen Schicht und bis 500 Neuronen schneiden alle ähnlich gut ab, wobei der größte Deckungsgrad bei 300 Neuronen und einem Wert von 0,0127 erreicht wird. Alle Konfigurationen mit mehreren Schichten liegen mit Abstand unter diesem Bestwert.

Interessanterweise fallen die Ergebnisse für die Testmenge B2 besser aus (Abbildung 24). Aber auch hier schneiden die Konfigurationen mit mehreren Schichten schlechter ab, als diejenigen mit nur einer Schicht und wenigen Neuronen. Das Maximum von 0,0185 wird bei 200 Neuronen in einer einzigen verdeckten Schicht erreicht. Die besseren Ergebnisse im Gegensatz zur Testmenge B1 lassen sich durch die nur halb so vielen verschiedenen Schlagwörtern erklären, die in der Testmenge enthalten sind. Zwar verfügt

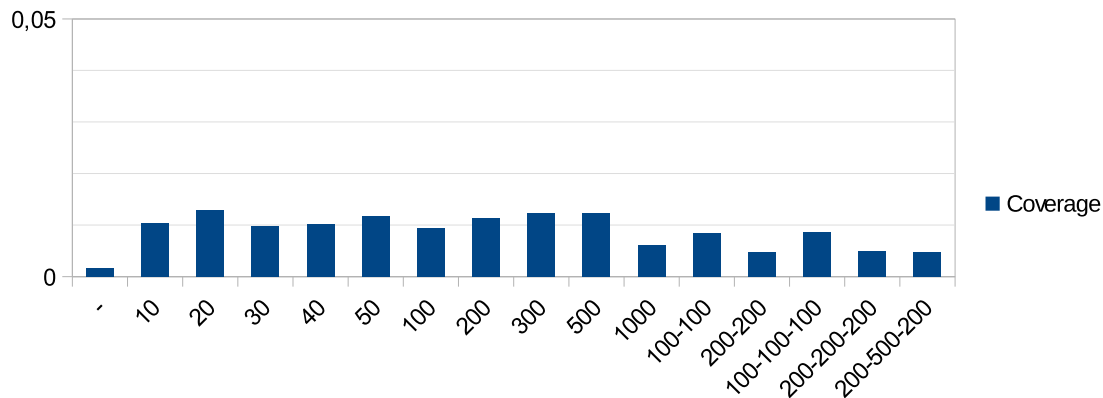


Abbildung 25: Ergebnisse Schlagwörter Netz Testmenge B3

die Testmenge B2 über weniger Datensätze für das Training, aber die geringere Anzahl Labels sorgt dennoch für eine kleine Verbesserung bei den Ergebnissen.

Bei der Testmenge B3 ist die Menge an Labels nochmals leicht größer als bei der Menge B1. Durch die gleichzeitig größere Menge an Datensätzen, die für das Training zur Verfügung stehen, fällt das Maximum des Deckungsgrades aber leicht höher aus (siehe Abbildung 25). So liegt es bei 0,0128 für eine Konfiguration mit einer verdeckten Schicht und 20 Neuronen. Die übrigen Charakteristika der Ergebnisse sind gleich zu der Testmenge B1.

Insgesamt gibt es für die Verschlagwortung mittels des neuronalen Netzes nicht, wie bei den Systemstellen, eine hervorstechende Konfiguration mit der die besten Ergebnisse erzielt werden. Bei jeder Testmenge wird für eine andere Konfiguration der Bestwert erreicht, allerdings liegen die Ergebnisse so dicht beieinander, dass kein deutliches Urteil der besten Konfiguration gefällt werden kann.

4.5 Anwendung auf Testdaten

An dieser Stelle werden abschließend nochmals die im vorigen Abschnitt bei den Kreuzvalidierungen ausgelassenen Teile der Testmengen A3 bzw. B3 ausgewertet. Die Parameter der Verfahren wurden auf Basis der Auswertungen in 4.3 und 4.4 getroffen und sind in Tabelle 9 dargestellt. Die Einzelergebnisse sind in Anhang F einzusehen.

Zunächst werden die Ergebnisse für die Systemstellen betrachtet. In Abbildung 26 sind die Ergebnisse der drei Verfahren für die Testmenge A3 abgebildet. Zur Vollständigkeit wurden auch für das neuronale Netz alle Maße angegeben. Aufgrund der Funktionsweise des Netzes ist aber nur der Coverage direkt vergleichbar.

	Systemstellen	Schlagwörter
BookGraph	t = 0,05 dyn	t = 0,05 dyn
KNN	k = 3	k = 2
Netz	500	20

Tabelle 9: Parameterwahl für die Anwendung auf den Testdaten

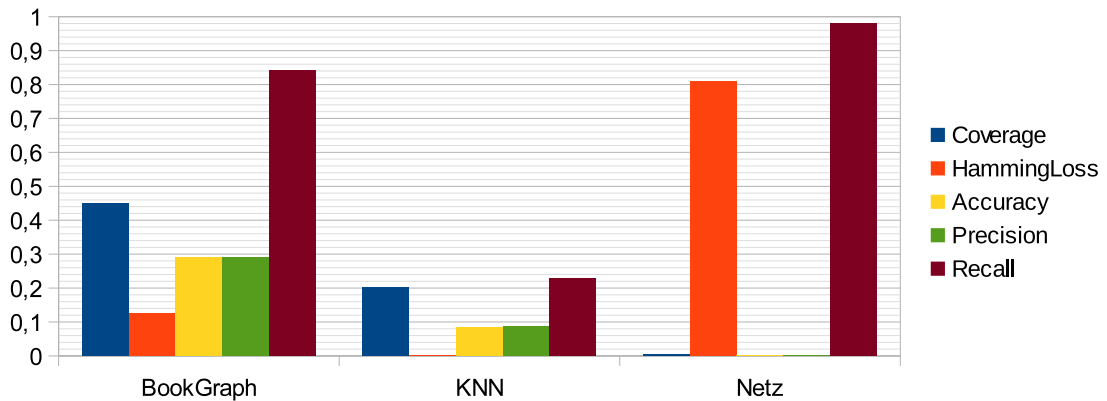
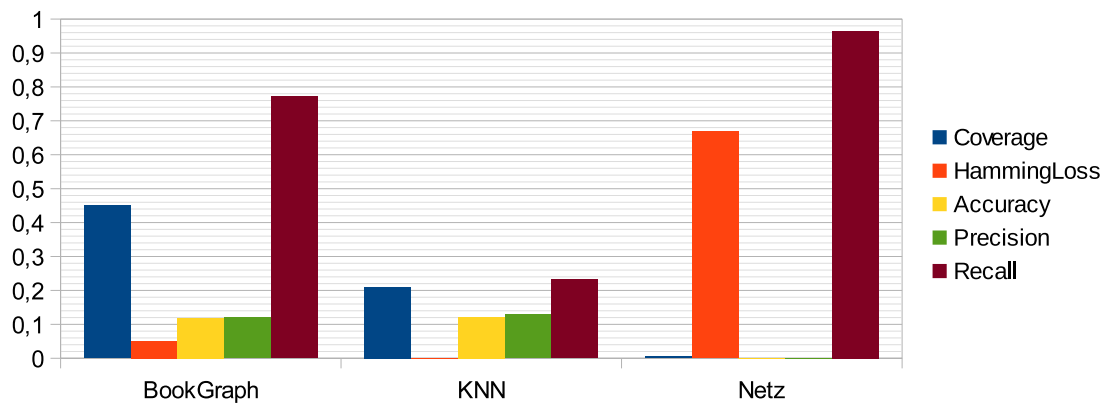


Abbildung 26: Ergebnisse der Testdaten für die Systemstellen

Das BookGraph Verfahren schneidet mit Abstand am besten ab. Es hat einen Deckungsgrad von 0,4493, während der KNN Klassifikator nur 0,2015 erreicht. Auch bei der Precision, der Accuracy und dem Recall liegt das BookGraph Verfahren deutlich vor dem KNN Klassifikator. Nur der HammingLoss fällt beim KNN Klassifikator geringer aus. Erstaunlicherweise liegen die Werte von BookGraph bei der Anwendung auf den Testdaten nochmals deutlich höher, als bei der Kreuzvalidierung. Dies könnte zum einen an der größeren Menge an Trainingsdaten liegen, aber auch zum anderen an zufälligerweise gut zu klassifizierenden Büchern in der Testmenge.

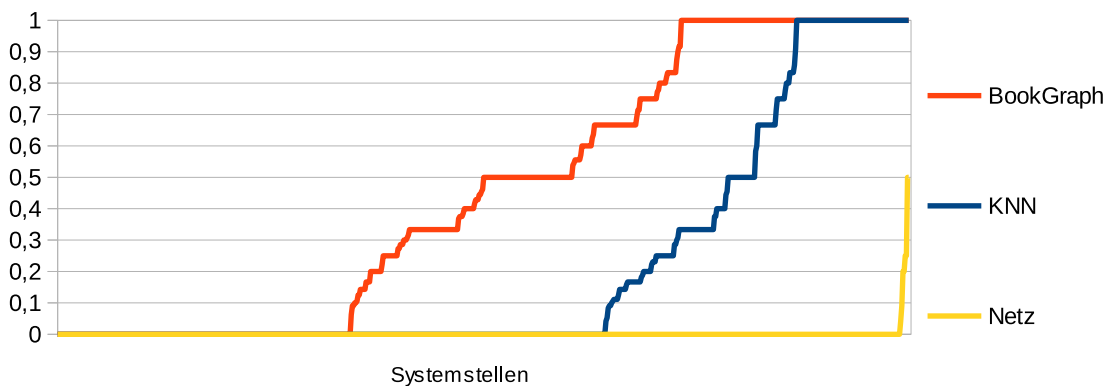
Im Gegensatz dazu kann das neuronale Netz in keiner Hinsicht überzeugen. Alle Werte sind durchweg sehr schlecht und der Recall ist nur durch die Arbeitsweise des Verfahrens so hoch.

Bei der Auswertung der Testdaten für die Schlagwörter ist auch das BookGraph Verfahren das beste, gefolgt vom KNN Klassifikator und dem neuronalen Netz als Schlusslicht. Während sich das BookGraph Verfahren klar bei Coverage und Recall absetzen kann, liegen Accuracy und Precision auf einer Ebene zum KNN.



4.6 Auswertung der Treffsicherheit

Die Treffsicherheit wird, analog zum Vorgehen in Abschnitt 4.5, für die Teile der Testmengen A3 und B3, die bei der Kreuzvalidierung ausgelassen wurden, ausgewertet. Anhand der Treffsicherheit kann analysiert werden, wie gut die Klassifikation für die einzelnen Labels funktioniert. Dafür wird für jedes Label der Recall bestimmt, dabei wird allerdings, wie beim Deckungsgrad, die jeweilige Vorschlagsliste auf die Länge an Gold-Standard Werten beschränkt.



In Abbildung 28 sind die prozentualen Ergebnisse für die Systemstellen dargestellt. Auf der x-Achse sind implizit für jedes Verfahren die Systemstellen aufsteigend eingetragen, während die y-Achse den Wert des jeweiligen Recalls angibt. Das neuronale Netz schafft es nur bei sehr wenigen Systemstellen diese zumindest einmal richtig vorzuschla-

gen. Die meisten Systemstellen werden somit niemals richtig vorgeschlagen und es gibt auch keine Systemstellen die 100% erreichen. Dahingegen liegt die Rate der gar nicht erfassten Systemstellen bei dem k -Nächste-Nachbarn Klassifikator bei rund 65%, d.h. ca. 35% werden zumindest teilweise richtigerweise vorgeschlagen. Davon erreichen über 10% einen Recall von 100%, werden also bei jedem Vorkommen in den Gold-Standard Werten auch vorgeschlagen. Das BookGraph Verfahren erreicht die höchsten Werte. Nur knapp 30% der Systemstellen werden niemals richtig vorgeschlagen, aber über 20% der Systemstellen immer.

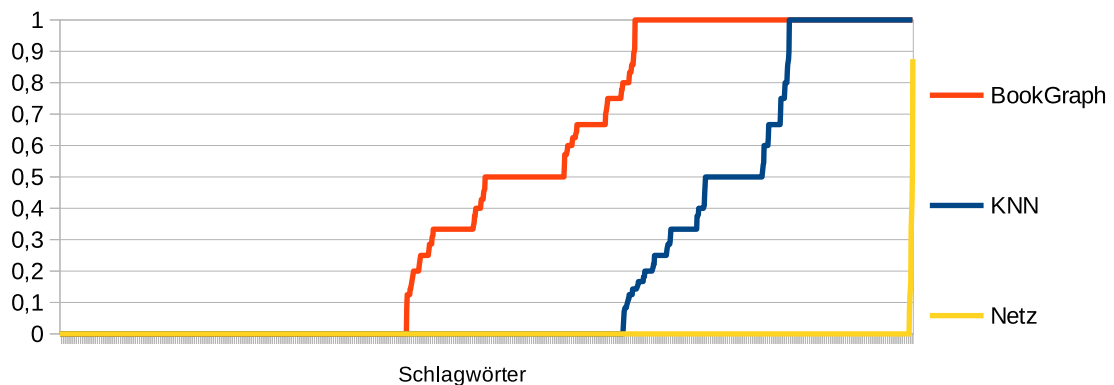


Abbildung 29: Ergebnisse der Treffsicherheit für die Schlagwörter

Für die Auswertung der Treffsicherheit bei den Schlagwörtern zeigt sich ein ähnliches Bild (siehe Abbildung 29). Das neuronale Netz findet wiederum nur sehr wenige richtige Vorschläge. So schlägt es nur 9 Schlagwörter richtig vor. Der KNN Klassifikator liegt bei ca. 14% der Schlagwörter immer richtig, erkennt allerdings auch 66% gar nicht. Auch hier schneidet das BookGraph Verfahren am besten ab. Über 32% der Schlagwörter werden zu 100% erkannt und nur knapp 40% gar nicht.

Insgesamt ist auffällig, dass sowohl bei den Systemstellen als auch bei den Schlagwörtern ein sehr hoher Anteil einen Recall von 0% aufweist. Dies sind vor allem die sehr selten auftretenden Systemstellen bzw. Schlagwörter, was mehrere Probleme verursacht. Zum einen existieren so nur wenige oder im Extremfall gar keine Trainingsdaten, die diese Systemstellen bzw. Schlagwörter aufweisen, und die Verfahren können diese somit nicht richtig erkennen. Zum anderen führt ein selten auftretendes Label bei der Auswertung zu sehr großen Sprüngen beim Wert. Eine Systemstelle, die nur ein einziges mal vorkommt, weist entweder einen Recall von 0% oder 100% auf, Zwischenwerte sind folglich nicht möglich.

4.7 Auswertung für Single-Label

Wie bereits beschrieben, werden in diesem Abschnitt diejenigen Bücher der Testmenge A3 (siehe 4.3) separat betrachtet, welche nur über eine einzige Systemstelle verfügen. Die resultierende Menge umfasst 90.739 Datensätze, welche wie zuvor mittels Kreuzvalidierung für jedes Verfahren ausgewertet werden. Dabei werden auch wieder verschiedene Werte für die Parameter der Verfahren getestet, um deren Einfluss auf das Ergebnis zu analysieren.

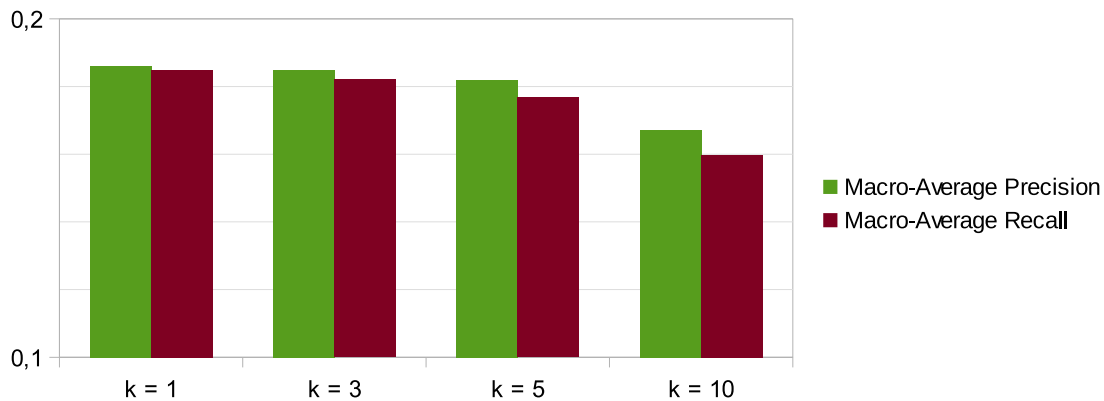


Abbildung 30: Ergebnisse KNN Single-Label

Von den Vorschlagslisten wird, wie in 4.2.5 beschrieben, nur jeweils der beste Vorschlag genommen, sodass diese mit den klassischen Maßen für Single-Label Klassifikationen ausgewertet werden können.

In Abbildung 30 sind die Ergebnisse für den k -Nächste-Nachbarn Klassifikator dargestellt. Der Fall $k = 2$ wird an dieser Stelle ausgelassen, da er bei gleichgewichteter Behandlung aller Nachbarn die selben Vorschläge macht wie für $k = 1$.

Für die Single-Label Auswertung zeigt sich im Gegensatz zur bisherigen Multi-Label Klassifikation ein interessantes Bild. Sowohl die Macro-Averaged Precision als auch der Macro-Averaged Recall steigen bei fallendem Wert für k . Dies bedeutet, dass für kleinere k sowohl die Qualität der Vorschläge steigt, als auch die Erkennungsrate der richtigen Systemstelle. Das Maximum für die Macro-Averaged Precision beträgt 0,1862 und für den Macro-Averaged Recall 0,1848.

Das BookGraph Verfahren kann im Vergleich zum KNN Klassifikator deutlich bessere Werte erzielen (siehe Abbildung 31). Auffällig ist, dass bei $t = 0,05$ und $t = 0,1$ es keine Auswirkung hat, ob der Schwellwert dynamisch ist oder nicht. Bei beiden Werten für den Schwellwert werden direkt genügend Vorschläge gefunden, sodass eine Absen-

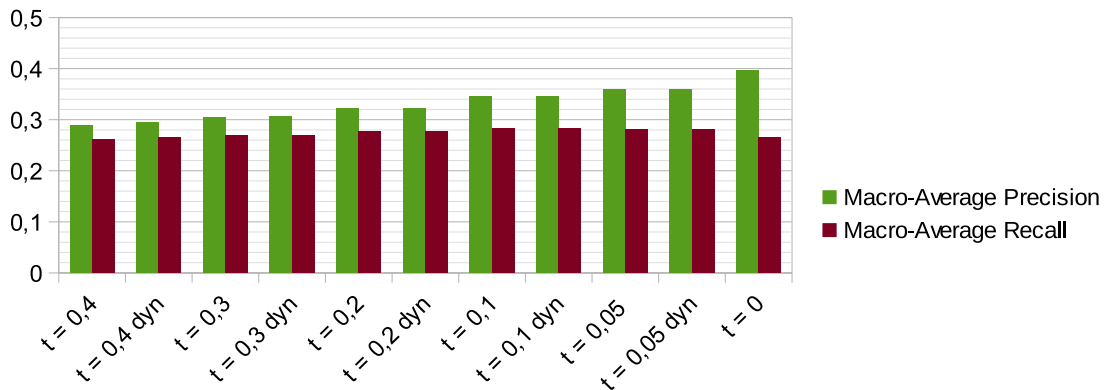


Abbildung 31: Ergebnisse BookGraph Single-Label

kung nicht notwendig ist. Erst bei einem Schwellwert von 0,2 wirkt sich die dynamische Schwelle aus und verbessert die Ergebnisse.

Das Weglassen des Schwellwerts bzw. Setzen des Schwellwerts auf 0 erzielt mit Abstand den besten Wert für die Macro-Averaged Precision von 0,3957. Bei $t = 0,05$ beträgt diese nur 0,3597. Der höchste Wert für den Macro-Averaged Recall von 0,2825 wird allerdings für $t = 0,1$ erzielt.

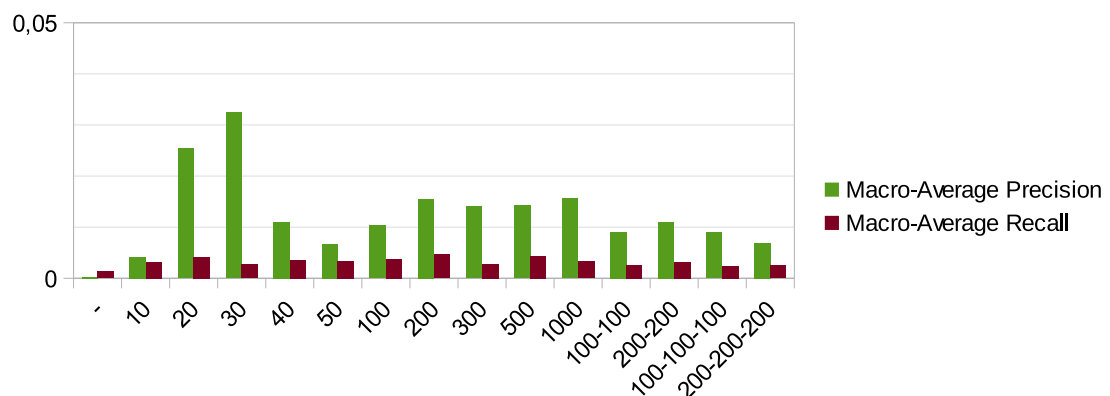


Abbildung 32: Ergebnisse Netz Single-Label

Auch bei der Single-Label Betrachtung schneidet das neuronale Netz schlechter als seine beiden Konkurrenten ab. In Abbildung 32 sind die Ergebnisse für verschiedene Konfigurationen der verdeckten Schichten aufgeführt. Für die Macro-Averaged Precision ist eine deutliche Spitze bei 20 bzw. 30 Neuronen in einer einzigen Schicht zu sehen. Für den Fall mit 30 Neuronen wird ein Maximum von 0,0324 erreicht.

Der Macro-Averaged Recall zeigt keine ausgeprägten Spitzen, sondern schwankt in

einem kleineren Bereich. Eine Konfiguration von 200 Neuronen in einer Schicht ergibt das Maximum mit 0,0047. Allerdings sind bei solch kleinen Differenzen zwischen den einzelnen Testläufen Aussagen zur besten Konfiguration nicht möglich, da die Varianz beim Training der neuronalen Netze für Schwankungen der Ergebnisse sorgt.

Alle Ergebniswerte sind der Arbeit angehängt (Anhang E).

4.8 Evaluation des Praxiseinsatzes durch Sachbearbeiter

In diesem Abschnitt der Evaluation soll die Praxistauglichkeit der Verfahren begutachtet werden. Dazu werden für bisher nicht sacherschlossene Bücher, aus ausgewählten Themengebieten, Vorschlagslisten für die Systemstellen und Schlagwörter erstellt und anschließend durch Sachbearbeiter der ZB bewertet. Die Bewertung erfolgt auf Grundlage dessen, welche Schlagwörter und Systemstellen der Sachbearbeiter von den Vorschlagslisten für sinnvoll erachtet und dem Buch für den Bibliothekskatalog zuweisen würde.

ID	Vorschläge	Wahrscheinlichkeit	Gütegrad
614229	FGN - Nanotechnology	0,1853	G1
	ETF - Fuel cells - general	0,1423	G1
	CNE - Electrochemistry	0,08431	G2
	CLM - Catalytic processes	0,0738	G1
	ESB - Hydrogen energy	0,0587	
	CLJ - Catalyst materials	0,0527	
374116	ETFG - Polymer electrolyte membrane fuel cells	0,5	G1
	ETFP - Fuel cell construction and modeling	0,5	G1

Tabelle 10: Beispiel für die Einteilung in Gütegrade der Systemstellen

Wie im Abschnitt 4.5 werden wieder die besten in dieser Arbeit ermittelten Parameter gewählt. Diese sind in der Tabelle 4.8 dargestellt.

Um die Testmengen zu erstellen wurden in Absprache mit den Sachbearbeitern Suchanfragen zu bestimmten Themen an das Bibliothekssystem gestellt. Aus den Treffern wurden dann nach dem Zufallsprinzip pro Suchanfrage zehn E-Bücher, welche noch keine Schlagwörter oder Systemstellen zugewiesen haben und über ein Inhaltsverzeichnis verfügen, ausgewählt. Die Suchanfragen sind „Fuel Cell“, „Energy Storage“ und „Spectroscopy“.

Die Parameter der Verfahren wurden auf Basis der Auswertungen in 4.3 und 4.4 getroffen und sind in Tabelle 9 dargestellt. Als Trainingsdaten werden die kompletten Testmengen A3 bzw. B3 genutzt.

Damit die Sachbearbeiter eine übersichtliche Liste an Vorschlägen für die Schlagwörter und Systemstellen bekommen, werden diese teilweise abgeschnitten. In Abstimmung mit den Sachbearbeitern werden die Vorschlagslisten auf Einträge mit mindestens 5% Relevanz beschränkt, wobei dennoch auf mindestens fünf Vorschläge bei den Systemstellen bzw. zehn Vorschläge für die Schlagwörter aufgefüllt wird. Wie im Abschnitt 4.1 zu sehen, beträgt bei den E-Büchern die Label Kardinalität für die Systemstellen 1,5247 und für die Schlagwörter 3,4078, sodass eine Beschränkung der Anzahl Vorschläge von fünf bzw. zehn bis maximal 20 eine ausreichende Menge darstellt.

Die Vorschläge werden von den Sachbearbeitern in drei Gütegraden eingeteilt. Gütegrad 1 (G1) umfasst Vorschläge die richtig sind und so auch von den Sachbearbeitern vergeben würden. Der zweite Gütegrad (G2) ist für Vorschläge, welche zwar sachlich richtig sind, aber einen allgemeineren Überbegriff beschreiben. Als letztes bekommen falsche Vorschläge den Gütegrad 3 (G3) zugeordnet. Im Folgenden werden nur Vorschläge der Grade G1 und G2 betrachtet, da es für die Sachbearbeiter entbehrlich ist, wie viele falsche Vorschläge gemacht werden, solange auch richtige Vorschläge vertreten sind. Ein Beispiel für dieses Bewertungssystem ist anhand von zwei Büchern in Tabelle 10 zu sehen. Für das erste Buch werden sechs Vorschläge für die Systemstellen gemacht, da alle über 5% Wahrscheinlichkeit besitzen. Von diesen Vorschlägen werden drei mit G1, ein mit G2 und zwei mit G3 bewertet. Für das zweite Beispielbuch werden insgesamt nur zwei Vorschläge gefunden, sodass eine Auffüllung auf mindestens fünf Vorschläge nicht möglich ist.

		BookGraph	KNN	Netz
Fuel Cell	$ Bücher_{G1} $	8	7	0
	$ Bücher_{G1} \cup Bücher_{G2} $	9	8	1
	$ Vorschläge_{G1} $	11	7	0
	$ Vorschläge_{G2} $	3	3	1
Energy Storage	$ Bücher_{G1} $	8	0	2
	$ Bücher_{G1} \cup Bücher_{G2} $	9	3	3
	$ Vorschläge_{G1} $	13	0	2
	$ Vorschläge_{G2} $	6	3	1
Spectroscopy	$ Bücher_{G1} $	3	0	0
	$ Bücher_{G1} \cup Bücher_{G2} $	4	1	0
	$ Vorschläge_{G1} $	4	0	0
	$ Vorschläge_{G2} $	1	1	0

Tabelle 11: Ergebnisse für die Systemstellen bei der Evaluation durch Sachbearbeiter

In Tabelle 11 sind die Ergebnisse für die Bewertung der Systemstellenvorschläge für

die drei Suchanfragen und allen Verfahren zu sehen. $Bücher_{G1}$ ist die Menge derjenigen Bücher, die mindestens einen Vorschlag der Güteklasse G1 besitzen. Analog sind in $Bücher_{G2}$ die Bücher mit mindestens einem G2 Vorschlag enthalten. $Vorschläge_{G1}$ bzw. $Vorschläge_{G2}$ beinhalten alle Vorschläge der Klassen G1 bzw. G2.

Jede Suchanfrage wird separat betrachtet und mittels vier Maße ausgewertet. Angegeben werden die Anzahl an Büchern mit G1 Vorschlägen, die Anzahl Bücher mit G1 oder G2 Vorschlägen, die Gesamtzahl an Vorschlägen der Klasse G1 sowie die die Gesamtzahl an G2 Vorschlägen. Alle Maße werden absolut angegeben, da vor allem für die Anzahl an Vorschläge kein Soll oder Maximum existiert und so keine sinnvolle Relation genannt werden kann.

Das BookGraph Verfahren kann zu allen drei Suchanfragen sinnvolle Vorschläge finden. Wobei deutlich erkennbar ist, dass für die beiden spezielleren Suchen „Fuel Cell“ und „Energy Storage“ wesentlich bessere Vorschläge gefunden werden, als für die allgemeinere Suche nach „Spectroscopy“. Für die ersten beiden Suchen werden jeweils zu acht von zehn Büchern passende Systemstellen vorgeschlagen, wobei jedes dieser acht Bücher durchschnittlich 1,375 G1 Vorschläge aufweist. Bei der dritten Suche verfügen nur noch vier Bücher über G1 oder G2 Vorschläge.

		BookGraph	KNN	Netz
Fuel Cell	$ Bücher_{G1} $	8	7	4
	$ Bücher_{G1} \cup Bücher_{G2} $	9	8	7
	$ Vorschläge_{G1} $	21	10	5
	$ Vorschläge_{G2} $	12	7	5
Energy Storage	$ Bücher_{G1} $	9	2	3
	$ Bücher_{G1} \cup Bücher_{G2} $	9	3	4
	$ Vorschläge_{G1} $	28	8	3
	$ Vorschläge_{G2} $	4	3	2
Spectroscopy	$ Bücher_{G1} $	4	1	1
	$ Bücher_{G1} \cup Bücher_{G2} $	5	0	3
	$ Vorschläge_{G1} $	6	1	1
	$ Vorschläge_{G2} $	4	0	2

Tabelle 12: Ergebnisse für die Schlagwörter bei der Evaluation durch Sachbearbeiter

Für die spezielle Suche nach „Fuel Cell“ funktioniert das KNN Verfahren ähnlich gut wie das BookGraph Verfahren, auch wenn nur für sieben statt acht Büchern G1 Vorschläge gefunden werden. Zudem ist die Gesamtzahl an Vorschlägen auch geringer. Bei den anderen beiden Suchen versagt der KNN Klassifikator hingegen. Zu keinem der Bücher wird auch nur ein G1 Vorschlag und nur sehr wenige G2 Vorschläge gefunden.

Das neuronale Netz schafft es nur für die Suche nach „Energy Storage“ für zwei Bücher G1 Vorschläge zu machen. Für die anderen beiden Suchanfragen werden keine brauchbaren Systemstellen vorgeschlagen.

Die Evaluation der dazugehörigen Vorschläge für die Schlagwörter in Tabelle 12 zeigt ein ähnliches Bild für das Abschneiden der einzelnen Verfahren. Nur das BookGraph Verfahren schafft es für die beiden Suchanfragen nach „Fuel Cell“ und „Energy Storage“ genügend G1 Vorschläge zu finden, sodass die Ergebnisse als vollständig brauchbar gewertet werden können. Die anderen beiden Verfahren finden zwar auch G1 Vorschläge, aber deren Anzahl reicht nicht aus, um damit eine sinnvolle Verschlagwortung durchzuführen.

Für die dritte, allgemeinere Suche nach „Spectroscopy“ versagen das KNN Verfahren sowie das neuronale Netz, während das BookGraph Verfahren noch für vier Bücher G1 Vorschläge findet.

5 Programm zur Demonstration der Verfahren

In diesem Kapitel wird kurz die Umsetzung der Verfahren aus Kapitel 3 in ein Programm erläutert. Das Programm diente während der Masterarbeit für Test- und Demonstrationszwecke innerhalb der ZB. Zudem wurden mit Hilfe dieses Programms die Werte für die Evaluation berechnet.

Entwickelt wird das Programm in C# und dem .NET-Framework 4.5. Für die Oberfläche kommt das Windows Presentation Framework (WPF) zum Einsatz.

5.1 Programmstrukturierung

Die Klassenstruktur basiert auf dem „Model View ViewModel“ [Mic16] (MVVM) Entwurfsmuster, welches einen ähnlichen Aufbau wie das Model-View-Controller Muster aufweist, allerdings den Controller durch ein ViewModel ersetzt (siehe Abbildung 33). Dadurch ist es möglich, die verschiedenen Klassifikatoren auf einfache Weise auszuprobieren, da sie auf Programmebene mit wenig Aufwand gegeneinander getauscht werden können. Auch die Bewertung mittels der in Kapitel 4 vorgestellten Evaluationsmaße ist durch das MVVM Muster effizient umzusetzen.

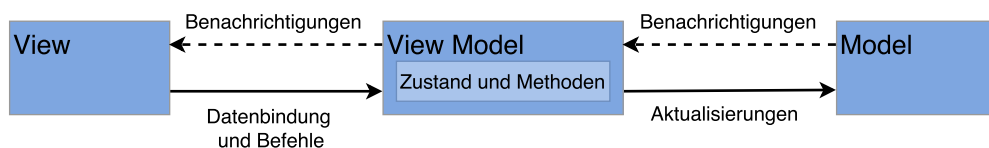


Abbildung 33: Das MVVM-Muster nach [Mic16]

5.2 Workflow

Zunächst wird mittels einer zweiten Konsolenanwendung eine Menge an Büchern von JuLib heruntergeladen und in eine Datei serialisiert. Die Auswahl der Bücher erfolgt direkt

als Suchanfrage an den Solr-Server von JuLib und funktioniert somit analog zur Weboberfläche des Bibliothekskatalog.

Die Bedienung des eigentlichen Programms erfolgt mittels einer grafischen Oberfläche. Der Benutzer lädt als erstes eine Menge von Büchern aus der zuvor erstellten Datei, um im weiteren Verlauf mit dieser Menge arbeiten zu können. Hat der Benutzer eine Auswahl an Büchern in das Programm geladen, kann er nun ein einzelnes Buch über eine Liste auswählen und sich so die Metadaten anzeigen lassen. Um eine automatische Sacherschließung des Buches durchführen zu können, muss zunächst ein Klassifikator erstellt werden. Dazu muss als erstes eine Auswahl für die Art des Klassifikators sowie die gewünschte Art der Vorschläge gewählt werden. Danach kann entweder ein neuer Klassifikator erzeugt oder es kann ein bereits vorhandener Klassifikator geladen werden. Über weitere Schaltflächen kann dann eine Vorschlagliste für Systemstellen oder Schlagwörter erstellt und angezeigt werden.

5.3 Grafische Benutzeroberfläche

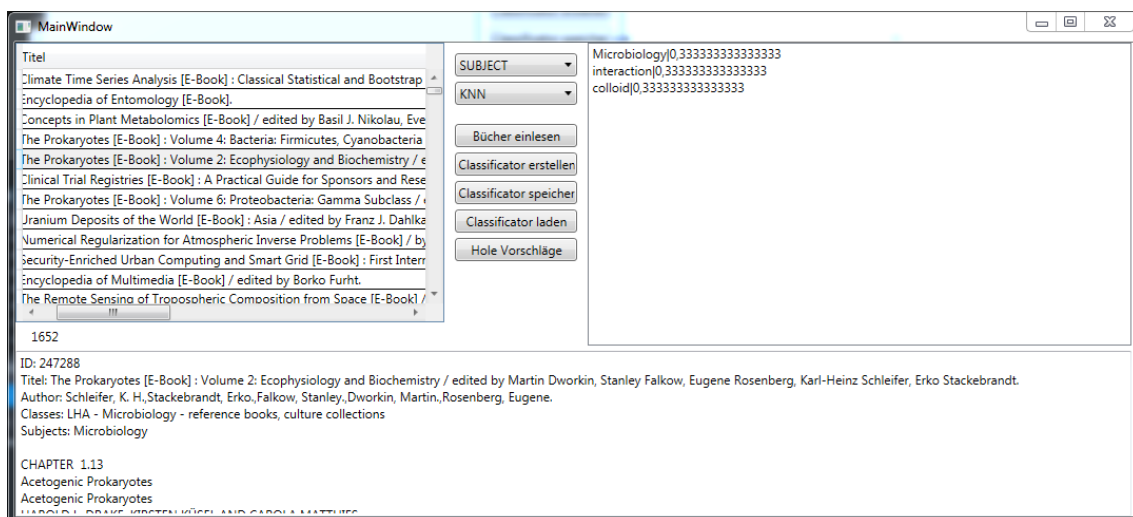


Abbildung 34: Oberfläche des Beispielprogramms

In Abbildung 34 ist die Oberfläche des Beispielprogramms zu sehen. Die wichtigsten Funktionen lassen sich über die zentralen Schaltflächen benutzen. Links daneben werden alle aktuell geladenen Bücher in einer Liste dargestellt. Wird ein Buch aus der Liste ausgewählt, werden im unteren Bereich der Oberfläche die weiteren Details angezeigt. Dazu gehören unter anderem Titel, Autoren, Sprache und, wenn vorhanden, Inhaltsverzeichnis, Schlagwörter und Systemstellen. Der letzte Bereich der Oberfläche ist eine Liste

an der rechten Seite, in der die Vorschlagsliste für die Systemstellen oder Schlagwörter angezeigt werden.

Über die zentralen Bedienelement kann ausgewählt werden, welcher Klassifikator genutzt werden soll und welche Art an Vorschlägen gewünscht ist. Zudem lässt sich ein Klassifikator neu erstellen, speichern und laden. Für die Sacherschließung von Büchern kann eine Menge an Büchern geladen und zu einem ausgewählten Buch eine Vorschlagsliste mit Systemstellen bzw. Schlagwörtern erstellt werden.

6 Fazit und Ausblick

Dieses Kapitel rekapituliert die Arbeit und fasst die Inhalte nochmals zusammen. Zum Schluss wird ein Ausblick gegeben.

6.1 Zusammenfassung und Fazit

In dieser Arbeit wurde eine mögliche automatische Verschlagwortung sowie Einordnung in die Systematik für die Bücher der Zentralbibliothek des Forschungszentrums Jülich angeschaut. Dafür wurden drei verschiedene Verfahren ausgewählt und im Verlauf der vorigen Kapitel auf ihre Leistungsfähigkeit hin untersucht.

BookGraph	KNN	Netz
+ erzielt besten Ergebnisse - sehr spezielles Verfahren - langwierige Erzeugung des Graphen nötig	+ annehmbare Ergebnisse + kein Training nötig - viele Trainingsdaten benötigt - anfällig für Ausreißer	+ sehr generischer Ansatz - sehr schlechte Ergebnisse - langwieriges Training - separate Netze für Schlagwörter und Systemstellen nötig

Tabelle 13: Vor- und Nachteile der Verfahren

Das BookGraph Verfahren wurde im Rahmen der Arbeit speziell für diesen Anwendungsfall entwickelt, während bei den anderen beiden Verfahren auf bereits bewährte Multi-Label Klassifikatoren zurückgegriffen wurde.

BookGraph erzeugt bei allen durchgeführten Tests die qualitativ besten Vorschläge. Der k -Nächste-Nachbarn Klassifikator liefert noch annehmbare Ergebnisse, wohingegen das neuronale Netz keine in der Praxis verwendbaren Vorschlagslisten erstellt. In Tabelle 13 sind die wesentlichen Vor- und Nachteile der Verfahren aufgezählt.

Müsste abschließend eines der drei Verfahren für einen Produktiveinsatz in der ZB ausgewählt werden, so wäre es das BookGraph Verfahren. Es liefert konstant die besten Vorschläge, sodass die wenigen Nachteile vernachlässigbar sind.

6.2 Ausblick

Zum Abschluss der Arbeit erfolgt in diesem Abschnitt ein kurzer Ausblick. Dabei wird zum einen die Weiterentwicklung der Verfahren beleuchtet und zum anderen die nötigen Schritte bis zum produktiven Einsatz in der Zentralbibliothek beschrieben.

6.2.1 Weiterentwicklung der Verfahren

Für eine Weiterentwicklung der Bestimmung von Systemstellen und Schlagwörtern gibt es mehrere mögliche Ansätze.

Zum einen könnten komplett andere Klassifikatoren für Multi-Label Probleme gewählt werden, sodass ein anderer Ansatz zur Anwendung kommt. Zum anderen könnten aber auch die in dieser Arbeit angesprochenen Verfahren modifiziert werden.

Die Feature-Vektoren, die vom KNN Klassifikator bzw. dem neuronalen Netz genutzt werden, bieten ein großes Potenzial für Änderungen. So könnten andere Feature gewählt werden, um eine Verbesserung der Ergebnisse zu erzielen, zum Beispiel könnten die einzelnen Word2Vec Vektoren anders genutzt werden, um daraus einen einzigen Vektor für ein Buch zu erzeugen. Auch am BookGraph sind Anpassungen für ein eventuell besseres Ergebnis möglich, indem andere Methoden zur Auswertung des Graphen entwickelt werden. Hier wäre eine Einbeziehung der Buchtitel oder Verlage denkbar, um daraus mögliche Vorschläge zu kreieren.

Eine Kombination der Verfahren wäre auch möglich. So könnten die Stärken beider Verfahren genutzt werden. Dabei ist nur zu überlegen, wie die Vorschläge der einzelnen Verfahren am besten kombiniert werden, um die bestmöglichen Vorschläge zu erhalten.

Außerdem kann das BookGraph Verfahren für abseits dieser Arbeit liegende Auswertungen angepasst werden. Beispielsweise könnte auch mit Hilfe des Graphen die Ähnlichkeit zwischen Schlagwörtern oder Autoren betrachtet werden. Des Weiteren ließe sich zusätzlich eine Bestimmung des vermutlichen Autors abbilden. Diese ist allerdings komplizierter und erwartungsgemäß ungenauer, da die Menge aller Autoren im Vergleich zu den Schlagwörtern oder Systemstellen keine abgeschlossene Menge ist und nur Autoren vorgeschlagen werden könnten, die bereits in den Trainingsdaten enthalten sind.

6.2.2 Produktiveinsatz

Da in dieser Arbeit die Verfahren zur maschinellen Bestimmung der Schlagwörter und Systemstellen im Vordergrund standen, sind für den produktiven Einsatz in der Zentralbibliothek weitere Arbeiten notwendig. Das in Kapitel 5 vorgestellte Beispielprogramm

kann zum aktuellen Stand nur Datensätze aus JuLib extrahieren und für diese Vorschlaglisten erzeugen und anzeigen. Damit das Programm den Anforderungen der Sachbearbeiter genügt, muss ein Editieren der Datensätze möglich sein. Ein Sachbearbeiter muss in der Lage sein aus der Vorschlagliste Einträge auszuwählen sowie fehlende Schlagwörter bzw. Systemstellen manuell einzufügen. Als letzten Schritt sollte das Programm daraufhin das Übertragen der geänderten Datensätze nach JuLib unterstützen.

Literaturverzeichnis

- [AL99] APPELRATH, Hans-Jürgen ; LUDEWIG, Jochen: *Skriptum Informatik - eine konventionelle Einführung*. 4. überarbeitete Auflage. 1999
http://www.iste.uni-stuttgart.de/fileadmin/user_upload/iste/se/research/publications/Skriptum_HJAx_JL_MEDOC/Skriptum_Informatik/Skriptum.html
- [CH06] COVER, T. ; HART, P.: Nearest Neighbor Pattern Classification. In: *IEEE Trans. Inf. Theor.* 13 (2006), September, Nr. 1, S. 21–27
- [Def16a] *Definiton Schlagwort*. <https://glossar.ub.uni-kl.de/begriff574>. Version: 2016. – Stand: 19.09.2016
- [Def16b] *Definiton Systematische Erschliessung*. <https://glossar.ub.uni-kl.de/begriff473>. Version: 2016. – Stand: 21.09.2016
- [Ebe11] EBERHARDT, Joachim: Was ist (bibliothekarische) Sacherschliessung? (2011)
- [GS04] GODBOLE, Shantanu ; SARAWAGI, Sunita: Discriminative Methods for Multi-labeled Classification. In: *Advances in Knowledge Discovery and Data Mining* (2004), S. 22–30
- [Kri07] KRIESEL, David: *Ein kleiner Überblick über Neuronale Netze*. 2007
- [Lib16] LIBRARY OF CONGRESS: *Cataloging and Acquisitions*. <http://www.loc.gov/aba/>. Version: 2016. – Stand: 20.11.2016
- [LML⁺99] LIBRARY OF CONGRESS. NETWORK DEVELOPMENT ; MARC STANDARDS OFFICE ; LIBRARY OF CONGRESS ; NATIONAL LIBRARY OF CANADA. STANDARDS AND SUPPORTS ; BRITISH LIBRARY. BIBLIOGRAPHIC STANDARDS AND SYSTEMS ; LIBRARY AND ARCHIVES CANADA. STANDARDS AND SUPPORT ; NATIONAL LIBRARY OF CANADA ; LIBRARY AND ARCHIVES CANADA: *MARC 21 Format for Bibliographic Data: Including Guidelines for Content Designation*. Library of Congress, Cataloging Distribution Service, 1999

- [MCCD13] MIKOLOV, Tomas ; CHEN, Kai ; CORRADO, Greg ; DEAN, Jeffrey: Efficient Estimation of Word Representations in Vector Space. In: *CoRR* abs/1301.3781 (2013)
- [Mic16] MICROSOFT: *The MVVM Pattern*. <https://msdn.microsoft.com/en-us/library/hh848246.aspx>. Version: 2016. – Stand: 21. November 2016
- [MSC⁺13] MIKOLOV, Tomas ; SUTSKEVER, Ilya ; CHEN, Kai ; CORRADO, Greg ; DEAN, Jeffrey: Distributed Representations of Words and Phrases and their Compositionality. In: *CoRR* abs/1310.4546 (2013)
- [RB93] RIEDMILLER, Martin ; BRAUN, Heinrich: A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In: *IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS*, 1993, S. 586–591
- [Ros58] ROSENBLATT, Frank: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. In: *Psychological Review* 65 (1958), Nr. 6, S. 386–408
- [ŘS10] ŘEHŮŘEK, Radim ; SOJKA, Petr: Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta : ELRA, Mai 2010, S. 45–50
- [Sou14] SOUZA, C. R.: The Accord.NET Framework. São Carlos, Brazil, Dec 2014
- [SS00] SCHAPIRE, Robert E. ; SINGER, Yoram: BoosTexter: A Boosting-based System for Text Categorization. In: *Mach. Learn.* 39 (2000), Mai, Nr. 2-3, S. 135–168
- [TK07] TSOUMAKAS, Grigorios ; KATAKIS, Ioannis: Multi-label classification: An overview. In: *Int J Data Warehousing and Mining* 2007 (2007), S. 1–13
- [Yan99] YANG, Yiming: An Evaluation of Statistical Approaches to Text Categorization. In: *Inf. Retr.* 1 (1999), Mai, Nr. 1-2, S. 69–90

Abbildungsverzeichnis

1	Auszug aus einem Graphen	12
2	Algorithmus zur Bestimmung der Systemstellen	14
3	Beispiel mit verschiedenen k	16
4	Beispiel eines neuronalen Netzes	18
5	Ergebnisse Systemstellen BookGraph Testmenge A1	26
6	Ergebnisse Systemstellen BookGraph Testmenge A2	27
7	Ergebnisse Systemstellen BookGraph Testmenge A3	28
8	Ergebnisse Systemstellen BookGraph Testmenge A4	28
9	Ergebnisse Systemstellen KNN Testmenge A1	29
10	Ergebnisse Systemstellen KNN Testmenge A2	30
11	Ergebnisse Systemstellen KNN Testmenge A3	30
12	Ergebnisse Systemstellen KNN Testmenge A4	31
13	Ergebnisse Systemstellen Netz Testmenge A1	32
14	Ergebnisse Systemstellen Netz Testmenge A2	32
15	Ergebnisse Systemstellen Netz Testmenge A3	33
16	Ergebnisse Systemstellen Netz Testmenge A4	34
17	Ergebnisse Schlagwörter BookGraph Testmenge B1	35
18	Ergebnisse Schlagwörter BookGraph Testmenge B2	36
19	Ergebnisse Schlagwörter BookGraph Testmenge B3	36
20	Ergebnisse Schlagwörter KNN Testmenge B1	37
21	Ergebnisse Schlagwörter KNN Testmenge B2	37
22	Ergebnisse Schlagwörter KNN Testmenge B3	38
23	Ergebnisse Schlagwörter Netz Testmenge B1	39
24	Ergebnisse Schlagwörter Netz Testmenge B2	39
25	Ergebnisse Schlagwörter Netz Testmenge B3	40
26	Ergebnisse der Testdaten für die Systemstellen	41
27	Ergebnisse der Testdaten für die Schlagwörter	42
28	Ergebnisse der Treffsicherheit für die Systemstellen	42

29	Ergebnisse der Treffsicherheit für die Schlagwörter	43
30	Ergebnisse KNN Single-Label	44
31	Ergebnisse BookGraph Single-Label	45
32	Ergebnisse Netz Single-Label	45
33	Das MVVM-Muster nach [Mic16]	51
34	Oberfläche des Beispielprogramms	52

Tabellenverzeichnis

1	Wichtigste Metadaten einer Publikation	6
2	Beispiel für ein Buch	7
3	Beispiel aus der Systematik	8
4	Übersicht über die Testdaten Verteilung	19
5	Werte für $LC(D)$ und $LD(D)$ des Testdatensatzes	20
6	Beispiel für die Evaluationsmaße	22
7	Testmengen für Systemstellen	25
8	Testmengen für Schlagwörter	34
9	Parameterwahl für die Anwendung auf den Testdaten	41
10	Beispiel für die Einteilung in Gütegrade der Systemstellen	46
11	Ergebnisse für die Systemstellen bei der Evaluation durch Sachbearbeiter	47
12	Ergebnisse für die Schlagwörter bei der Evaluation durch Sachbearbeiter .	48
13	Vor- und Nachteile der Verfahren	55

A Auszug der Systemstellen

010 - Simulation Science (textbook)

011 - Simulation Science (advanced reading)

A - General (journals)

AAA - Encyclopedias - general aspects

AAD - General handbooks of knowledge

AAE - Popular science - general aspects

ABA - General international and supranational biographic catalogues

ABD - National biographic catalogues - Germany

ABG - National biographic catalogues - Europe

ABH - National biographic catalogues - non-Europe

ABM - Special subject biographies

ACA - Languages, dictionaries - general

ACE - Germanic languages, Indo-Germanic languages

ACH - Romance languages

ACM - East european languages

ACQ - Oriental languages

ACZ - Lists of abbreviations

AEA - Corporations - official - university - etc

AED - Official corporations - international and supranational

AEG - Official corporations - Germany

AEK - Official corporations - Europe

AEM - Official corporations - non-Europe

AEQ - Universities, academies - international and supranational

AES - Universities, academies - Germany

AET - Universities, academies - Europe

AEX - Universities, academies - non-Europe

AFA - Research institutions, scientific organizations - international and supranational

AFC - Research institutions, scientific organizations - Germany (DFG, MPG, FHG)

B Auszug der Schlagwörter

EMG (electromyography)	Laser pulses, Ultrashort
ENDOR (electron nuclear double resonance)	Lebesgue Stieltjes integral
ESR (electron spin resonance)	Lebesgue integral
ESR spectrometer	Legendre function
ESR spectroscopy	Legendre transform
ESR spectrum	Lewis acid
EXAFS (X-ray absorption fine structure)	Lie algebra
Eclipse	Lie group
Eiffel (programming language)	Lie series
Elbe river	Logo (programming language)
Electronic information resources	Lorentz equation
Endocrinology	Lorentz transform
Escherichia coli	Lorenz attractor
European Union	Low Carbon Economy (LCE)
European single market	Lurgi process
Europe	Lyapunov exponent
Eva Adam system	Lyapunov function
Evolution	Lyapunov method
Exxon gasification process	Lyapunov stability
FLR (field line resonance) frequency	MEG (magnetoencephalography)
FMRB reactor	MHD cell
FORTRAN (programming language)	MHD generator
FORTRAN 2008 (programming language)	MHD power plant
FORTRAN 77 (programming language)	MHD power
FORTRAN 90 (programming language)	MHD wave
FORTRAN 95, F (programming language)	MIMD computer
FRJ 1 reactor	MIS transistor
FRJ 2 reactor	MODULA (programming language)

C Einzelergebnisse Systemstellen

Auswertung

Klassifikator	Testmenge	Parameter	Coverage	HammingLoss	Accuracy	Precision	Recall
BookGraph	A1	t = 0,4	0,2576	0,0081	0,1202	0,1318	0,4408
		t = 0,4 dyn	0,2713	0,0091	0,1094	0,1164	0,4926
		t = 0,3	0,2849	0,0127	0,1179	0,1287	0,5186
		t = 0,3 dyn	0,2910	0,0132	0,1090	0,1159	0,5453
		t = 0,2	0,3047	0,0194	0,1132	0,1230	0,5915
		t = 0,2 dyn	0,3080	0,0196	0,1053	0,1120	0,6044
		t = 0,1	0,3282	0,0450	0,0999	0,1073	0,7022
		t = 0,1 dyn	0,3288	0,0451	0,0958	0,1022	0,7055
		t = 0,05	0,3342	0,0901	0,0914	0,0979	0,7794
		t = 0,05 dyn	0,3342	0,0901	0,0903	0,0966	0,7803
		t = 0	0,3030	0,6231	0,0848	0,0911	0,9005
	A2	t = 0,4	0,1341	0,0168	0,0618	0,0639	0,3417
		t = 0,4 dyn	0,1388	0,0174	0,0593	0,0608	0,3570
		t = 0,3	0,1419	0,0256	0,0569	0,0582	0,4102
		t = 0,3 dyn	0,1436	0,0258	0,0564	0,0578	0,4127
		t = 0,2	0,1490	0,0370	0,0539	0,05748	0,4705
		t = 0,2 dyn	0,1505	0,0371	0,0549	0,0561	0,4756
		t = 0,1	0,1481	0,0713	0,0485	0,0501	0,5704
		t = 0,1 dyn	0,1481	0,0713	0,0486	0,0496	0,5724
		t = 0,05	0,1391	0,1142	0,0455	0,0467	0,6303
		t = 0,05 dyn	0,1391	0,1142	0,0454	0,0464	0,6308
		t = 0	0,1145	0,3445	0,0422	0,0432	0,7776
	A3	t = 0,4	0,2576	0,0081	0,1202	0,1318	0,4408
		t = 0,4 dyn	0,2880	0,0086	0,1252	0,1313	0,5049
		t = 0,3	0,2987	0,0118	0,1360	0,1476	0,5260

Klassifikator	Testmenge	Parameter	Coverage	HammingLoss	Accuracy	Precision	Recall
BookGraph	A3	t = 0,3 dyn	0,3068	0,0124	0,1255	0,1316	0,5562
		t = 0,2	0,3194	0,0183	0,1310	0,1406	0,6006
		t = 0,2 dyn	0,3229	0,0185	0,1219	0,1278	0,6157
		t = 0,1	0,3395	0,0425	0,1167	0,1234	0,7097
		t = 0,1 dyn	0,3398	0,0426	0,1122	0,1177	0,7141
		t = 0,05	0,3465	0,0855	0,1079	0,1136	0,7871
		t = 0,05 dyn	0,3465	0,0855	0,1066	0,1121	0,7882
		t = 0	0,3131	0,6186	0,1010	0,1065	0,9036
	A4	t = 0,4	0,7473	0,0033	0,2937	0,3039	0,9466
		t = 0,4 dyn	0,7400	0,0040	0,2554	0,2620	0,9665
		t = 0,3	0,7426	0,0034	0,2705	0,2791	0,9601
		t = 0,3 dyn	0,7376	0,0043	0,2512	0,2578	0,9682
		t = 0,2	0,7019	0,0047	0,2542	0,2619	0,9681
		t = 0,2 dyn	0,7008	0,0047	0,2442	0,2507	0,9707
		t = 0,1	0,5867	0,0052	0,2400	0,2468	0,9722
		t = 0,1 dyn	0,5869	0,0052	0,2367	0,2432	0,9725
		t = 0,05	0,6137	0,0055	0,2354	0,2419	0,9732
		t = 0,05 dyn	0,6137	0,0055	0,2341	0,2406	0,9733
KNN	A1	k = 1	0,1804	0,0013	0,1748	0,2048	0,2041
		k = 2	0,1879	0,0020	0,1226	0,1359	0,2394
		k = 3	0,1892	0,0027	0,0906	0,0980	0,2590
		k = 5	0,1854	0,0042	0,0587	0,0619	0,2860
		k = 10	0,1677	0,0080	0,0322	0,0331	0,3177
	A2	k = 1	0,0893	0,0017	0,0819	0,1009	0,1009
		k = 2	0,0941	0,0024	0,0557	0,0641	0,1182
		k = 3	0,0914	0,0032	0,0442	0,0491	0,1362
		k = 5	0,0859	0,0047	0,0308	0,0332	0,1508
		k = 10	0,0733	0,0085	0,0190	0,0198	0,1772
	A3	k = 1	0,1985	0,0013	0,1933	0,2231	0,2219
		k = 2	0,2063	0,0020	0,1312	0,1448	0,2567
		k = 3	0,2081	0,0027	0,0963	0,1038	0,2772
		k = 5	0,2047	0,0042	0,0621	0,0652	0,3019
		k = 10	0,1827	0,0080	0,0343	0,0353	0,3370

Klassifikator	Testmenge	Parameter	Coverage	HammingLoss	Accuracy	Precision	Recall
KNN	A4A4	k = 1	0,5858	0,0005	0,5805	0,6389	0,6357
		k = 2	0,6139	0,0008	0,4663	0,4970	0,7199
		k = 3	0,6067	0,0011	0,3722	0,3893	0,7673
		k = 5	0,5861	0,0018	0,2610	0,2679	0,8282
		k = 10	0,5629	0,0032	0,1641	0,1657	0,9044
Netz	A1	-	0,0023				
		10	0,0142				
		20	0,0160				
		30	0,0176				
		40	0,0144				
		50	0,0167				
		100	0,0135				
		200	0,0121				
		300	0,0095				
		500	0,0210				
		1000	0,0042				
		100-100	0,0122				
		200-200	0,0111				
		100-100-100	0,0111				
		200-200-200	0,0076				
		200-500-200	0,0097				
	A2	-	0,0051				
		10	0,0077				
		20	0,0105				
		30	0,0087				
		40	0,0107				
		50	0,0147				
		100	0,0107				
		200	0,0072				
		300	0,0050				
		500	0,0118				
		1000	0,0021				
		100-100	0,0113				

Klassifikator	Testmenge	Parameter	Coverage	HammingLoss	Accuracy	Precision	Recall
Netz	A2	200-200	0,0092				
		100-100-100	0,0094				
		200-200-200	0,0065				
		200-500-200	0,0107				
	A3	-	0,0019				
		10	0,0131				
		20	0,0124				
		30	0,0147				
		40	0,0138				
		50	0,0149				
		100	0,0157				
		200	0,0144				
		300	0,0109				
		500	0,0185				
		1000	0,0065				
		100-100	0,0084				
		200-200	0,0118				
		100-100-100	0,0119				
		200-200-200	0,0099				
		200-500-200	0,0099				
	A4	-	0,1163				
		10	0,1907				
		20	0,2127				
		30	0,2300				
		40	0,2086				
		50	0,2397				
		100	0,2996				
		200	0,3484				
		300	0,3950				
		400	0,3199				
		500	0,2159				
		1000	0,1482				
		100-100	0,2694				

Klassifikator	Testmenge	Parameter	Coverage	HammingLoss	Accuracy	Precision	Recall
Netz	A4	200-200	0,2975				
		100-100-100	0,1525				
		200-200-200	0,2746				
		200-200-200-200	0,1427				

D Einzelergebnisse Schlagwörter

Auswertung

Klassifikator	Testmenge	Parameter	Coverage	HammingLoss	Accuracy	Precision	Recall
BookGraph	B1	t = 0,4	0,1341	0,0016	0,0696	0,0913	0,2068
		t = 0,4 dyn	0,1753	0,0024	0,0473	0,0535	0,3101
		t = 0,3	0,1620	0,0026	0,0794	0,1000	0,2825
		t = 0,3 dyn	0,1835	0,0032	0,0459	0,0514	0,3424
		t = 0,2	0,1895	0,0042	0,0818	0,1002	0,3575
		t = 0,2 dyn	0,2008	0,0046	0,0442	0,0487	0,3948
		t = 0,1	0,2350	0,0118	0,0730	0,0878	0,5245
		t = 0,1 dyn	0,2397	0,0130	0,0350	0,0373	0,5503
		t = 0,05	0,2646	0,0288	0,0574	0,0666	0,6519
		t = 0,05 dyn	0,2658	0,0291	0,0327	0,0349	0,6660
	B2	t = 0,4	0,0792	0,0032	0,0388	0,0476	0,1448
		t = 0,4 dyn	0,0966	0,0037	0,0309	0,0350	0,1900
		t = 0,3	0,1025	0,0054	0,0470	0,0555	0,2138
		t = 0,3 dyn	0,1095	0,0057	0,0318	0,0356	0,2352
		t = 0,2	0,1159	0,0091	0,0479	0,0588	0,2806
		t = 0,2 dyn	0,1183	0,0093	0,0302	0,0333	0,2916
		t = 0,1	0,1330	0,0272	0,0395	0,0447	0,4299
		t = 0,1 dyn	0,1342	0,0273	0,0246	0,0265	0,4379
		t = 0,05	0,1384	0,0576	0,0306	0,0341	0,5366
		t = 0,05 dyn	0,1388	0,0577	0,0214	0,0229	0,5410
	B3	t = 0,4	0,1376	0,0015	0,0754	0,0984	0,2091
		t = 0,4 dyn	0,1948	0,0024	0,0544	0,0611	0,3305
		t = 0,3	0,1738	0,0025	0,0890	0,1099	0,2905
		t = 0,3 dyn	0,2047	0,0031	0,0530	0,0590	0,3627
		t = 0,2	0,2033	0,0040	0,0935	0,1149	0,3659

Klassifikator	Testmenge	Parameter	Coverage	HammingLoss	Accuracy	Precision	Recall
BookGraph	B3	t = 0,2 dyn	0,2202	0,0044	0,0509	0,0561	0,4110
		t = 0,1	0,2511	0,0114	0,0864	0,1015	0,5294
		t = 0,1 dyn	0,2565	0,0130	0,0399	0,0423	0,5604
		t = 0,05	0,2813	0,0276	0,0711	0,0809	0,6601
		t = 0,05 dyn	0,2823	0,0279	0,0410	0,0435	0,6758
KNN	B1	k = 1	0,1318	0,0008	0,1210	0,1636	0,1651
		k = 2	0,1402	0,0012	0,0813	0,0993	0,1923
		k = 3	0,1375	0,0016	0,0590	0,0686	0,2070
		k = 5	0,1240	0,0025	0,0390	0,0432	0,2275
		k = 10	0,1007	0,0046	0,0224	0,0238	0,2554
	B2	k = 1	0,0666	0,0009	0,0600	0,0809	0,0829
		k = 2	0,0715	0,0013	0,0407	0,0513	0,0995
		k = 3	0,0673	0,0018	0,0318	0,0379	0,1145
		k = 5	0,0615	0,0026	0,0225	0,0254	0,1301
		k = 10	0,0480	0,0047	0,0140	0,0149	0,1502
	B3	k = 1	0,1476	0,0008	0,1375	0,1800	0,1809
		k = 2	0,1555	0,0012	0,0895	0,1076	0,2088
		k = 3	0,1533	0,0016	0,0655	0,0752	0,2241
		k = 5	0,1351	0,0025	0,0429	0,0471	0,2454
		k = 10	0,1081	0,0047	0,0243	0,0257	0,2744
Netz	B1	-	0,0026				
		10	0,0122				
		20	0,0105				
		30	0,0101				
		40	0,0111				
		50	0,0111				
		100	0,0108				
		200	0,0116				
		300	0,0127				
		500	0,0115				
		1000	0,0055				
		100-100	0,0082				
		200-200	0,0037				
		100-100-100	0,0065				

Klassifikator	Testmenge	Parameter	Coverage	HammingLoss	Accuracy	Precision	Recall
Netz	B1	200-200-200	0,0045				
		200-500-200	0,0079				
	B2	-	0,0046				
		10	0,0162				
		20	0,0118				
		30	0,0138				
		40	0,0159				
		50	0,0113				
		100	0,0154				
		200	0,0185				
		300	0,0082				
		500	0,0122				
		1000	0,0037				
		100-100	0,0133				
		200-200	0,0081				
		100-100-100	0,0108				
		200-200-200	0,0085				
		200-500-200	0,0096				
	B3	-	0,0017				
		10	0,0104				
		20	0,0128				
		30	0,0098				
		40	0,0101				
		50	0,0118				
		100	0,0095				
		200	0,0113				
		300	0,0123				
		500	0,0122				
		1000	0,0061				
		100-100	0,0085				
		200-200	0,0048				
		100-100-100	0,0085				
		200-200-200	0,0050				
		200-500-200	0,0048				

E Einzelergebnisse Single-Label Auswertung

Klassifikator	Parameter	Macro-Averaged Precision	Macro-Averaged Recall
BookGraph	t = 0,4	0,2893	0,2623
	t = 0,4 dyn	0,2949	0,2654
	t = 0,3	0,3038	0,2688
	t = 0,3 dyn	0,3056	0,2697
	t = 0,2	0,3212	0,2760
	t = 0,2 dyn	0,3222	0,2769
	t = 0,1	0,3453	0,2825
	t = 0,1 dyn	0,3453	0,2825
	t = 0,05	0,3597	0,2808
	t = 0,05 dyn	0,3597	0,2808
	t = 0	0,3957	0,2649
KNN	k = 1	0,1862	0,1848
	k = 3	0,1850	0,1821
	k = 5	0,1819	0,1767
	k = 10	0,1671	0,1598
Netz	-	0,0002	0,0014
	10	0,0042	0,0032
	20	0,0254	0,0042
	30	0,0324	0,0027
	40	0,0110	0,0036
	50	0,0067	0,0034
	100	0,0103	0,0038
	200	0,0155	0,0047
	300	0,0141	0,0028
	500	0,0143	0,0043

Klassifikator	Parameter	Macro-Averaged Precision	Macro-Averaged Recall
Netz	1000	0,0156	0,0033
	100-100	0,0090	0,0026
	200-200	0,0110	0,0031
	100-100-100	0,0090	0,0024
	200-200-200	0,0069	0,0025

F Einzelergebnisse Testdaten

Testmenge	Klassifikator	Coverage	HammingLoss	Accuracy	Precision	Recall
A1	BookGraph	0,3554	0,0868	0,1164	0,1201	0,8097
	KNN	0,2093	0,0028	0,0924	0,0995	0,2693
	Netz	0,0233	0,7925	0,0011	0,0011	0,9855
A2	BookGraph	0,1954	0,1096	0,0649	0,0664	0,6065
	KNN	0,0962	0,0030	0,0505	0,0544	0,1409
	Netz	0,0169	0,4056	0,0015	0,0015	0,7887
A3	BookGraph	0,4493	0,1259	0,2892	0,2899	0,8404
	KNN	0,2015	0,0026	0,0841	0,0866	0,2286
	Netz	0,0041	0,8086	0,0010	0,0010	0,9807
B1	BookGraph	0,2878	0,0277	0,0458	0,0494	0,6896
	KNN	0,1493	0,0012	0,0855	0,1056	0,1979
	Netz	0,0143	0,6441	0,0007	0,0007	0,9700
B2	BookGraph	0,1344	0,0559	0,0489	0,0500	0,5586
	KNN	0,0991	0,0012	0,0476	0,0585	0,1169
	Netz	0,0115	0,2811	0,0010	0,0010	0,7520
B3	BookGraph	0,4499	0,0511	0,1190	0,1202	0,7719
	KNN	0,2099	0,0011	0,1215	0,1292	0,2324
	Netz	0,0051	0,6685	0,0007	0,0007	0,9647